THIS WEEK

EDITORIALS

WORLD VIEW How London kept people away during the Olympics **p.337**

BIRD BRAIN Rattling boxes show how parrots can infer hidden objects **p.338**

RUBBING ROCKS Stone made smooth by 40,000 hours of earthquakes **p.339**

Through the gaps

A 20-year campaign of scientific fraud says as much about the research community as it does about the perpetrator. The system that allowed such deception to continue must be reformed.

Many questions are provoked by the shocking case of Yoshitaka Fujii, the Japanese anaesthesiologist who seems likely to set a record for the highest number of retracted papers by a single scientist. His entire list of publications has come under scrutiny: his trail of deception seems to have wound through almost 200 scientific articles over 20 years. Twenty years! How could it go on for so long?

As the News story on page 346 details, Fujii seems to have fabricated multiple studies wholesale, in some cases inventing participants. Nobody noticed — not his collaborators, funders, home institutions or journal editors. Or at least, nobody took action.

In retrospect, as in all cases of scientific fraud, the bulk of the questions will, rightly, focus on how to make sure that it cannot happen again. That, and why so much time passed before anyone investigated how Fujii was publishing clinical studies at impossible speed.

Fujii pulled the wool over the eyes of many different people — chief among them, various employers, whom he also falsely claimed had approved his studies, and journal editors. (One editor has publicly issued a mea culpa.) Perhaps most puzzling is that Fujii fooled his co-authors, one of whom published dozens of papers with him. The co-authors say that they had no suspicions; the Japanese Society of Anesthesiologists, which had a key role in exposing Fujii's fraud, is investigating.

But let's be honest. Even assuming that any co-author had suspicions, the current system means that it would not have been easy to raise the alert. It can be difficult to document a colleague's errant ways, and whistle-blowers might put their own careers at risk by angering a senior member of the field.

Those who inform authorities about other types of fraud sometimes get rewards. For example, the US government last week paid out its — and probably the world's — biggest ever payment to a whistle-blower. The former banker, who was jailed for his own role in a tax-evasion scandal, received US$104 million. Observers — especially lawyers — are pointing out that such windfalls might be the only way to encourage more insiders to put their necks on the line, which remains the most effective way to protect against such crimes.

That method is probably unworkable in science. Funders won't have that kind of cash to throw at scientific whistle-blowers. And imagine the uproar, not least in these pages, if whistle-blowers routinely got payouts bigger than the grants available for science projects through competitive peer review.

In the tax-evasion case, the figure was justified because it was only a small fraction of what the US government was able to recoup. But governments should also consider the amount of waste incurred by research fraud, especially when that fraud is carried out over decades and enshrouded in the scientific literature. On financial grounds alone, there are sound reasons for the authorities to increase the resources invested in efforts to limit academic misconduct, without the need to provide monetary rewards.

Japan, for example, could make it easier for whistle-blowers to take their claims to an external body, rather than to their employers. In theory, the country already has such a system. But in practice, agencies at the relevant ministries merely forward claims to the institutions involved, leaving whistle-blowers vulnerable.

In the wake of the latest scandal, there are signs of positive change. The Japanese Society of Anesthesiologists was so frustrated at the lack of an effective whistle-blowing mechanism that it plans to establish one. A group of 23 journal editors deserves credit for effectively, if belatedly, rooting out Fujii's problematic publications. And statistical approaches to evaluating results — such as those used to show that Fujii's data were far too perfect — are becoming more familiar, more readily available and, hopefully, more accepted as a legitimate way to audit published findings and raise red flags where necessary.

"On financial grounds alone, there are sound reasons for the authorities to increase the resources invested in efforts to limit academic misconduct."

It is important to note that although this latest case of fraud seems (again) to be an anomalous, extreme example involving one individual, the problems that allowed it to persist are endemic in scientific communities around the world. It is equally important to say (again) that they must be addressed in comprehensive fashion. ■

Extreme weather

Better models are needed before exceptional events can be reliably linked to global warming.

As climate change proceeds — which the record summer melt of Arctic sea-ice suggests it is doing at a worrying pace — nations, communities and individual citizens may begin to seek compensation for losses and damage arising from global warming. Climate scientists should be prepared for their skills one day to be probed in court. Whether there is a legal basis for such claims, such as that brought against the energy company ExxonMobil by the remote Alaskan community of Kivalina, which is facing coastal erosion and flooding as the sea ice retreats, is far from certain, however. So lawyers, insurers and climate negotiators are watching with interest the emerging ability, arising from improvements in climate models, to calculate how anthropogenic global warming will change, or has changed, the probability and magnitude of extreme weather and other climate-related events. But to make this emerging science of 'climate attribution' fit to inform legal and societal decisions will require enormous research effort.

Attribution is the attempt to deconstruct the causes of observable weather and to understand the physics of why extremes such as floods and heatwaves occur. This is important basic research. Extreme weather and changing weather patterns — the obvious manifestations of global climate change — do not simply reflect easily identifiable changes in Earth's energy balance such as a rise in atmospheric temperature. They usually have complex causes, involving anomalies in atmospheric circulation, levels of soil moisture and the like. Solid understanding of these factors is crucial if researchers are to improve the performance of, and confidence in, the climate models on which event attribution and longer-term climate projections depend.

Event attribution is one of the proposed 'climate services' — seasonal climate prediction is another — that are intended to provide society with the information needed to manage the risks and costs associated with climate change. Advocates of climate services see them as a counterpart to the daily weather forecast. But without the computing capacity of a well-equipped national meteorological office, heavily model-dependent services such as event attribution and seasonal prediction are unlikely to be as reliable.

At a workshop last week in Oxford, UK, convened by the Attribution of Climate-related Events group — a loose coalition of scientists from both sides of the Atlantic — some speakers questioned whether event attribution was possible at all. It currently rests on a comparison of the probability of an observed weather event in the real world with that of the 'same' event in a hypothetical world without global warming. One critic argued that, given the insufficient observational data and the coarse and mathematically far-from-perfect climate models used to generate attribution claims, they are unjustifiably speculative, basically unverifiable and better not made at all. And even if event attribution were reliable, another speaker added, the notion that it is useful for any section of society is unproven.

Both critics have a point, but their pessimistic conclusion — that climate attribution is a non-starter — is too harsh. It is true that many climate models are currently not fit for that purpose, but they can be improved. Evaluation of how often a climate model produces a good representation of the type of event in question, and whether it does so for the right reasons, must become integral to any attribution exercise. And when communicating their results, scientists must be open about shortcomings in the models used.

> ***"To make this emerging science of 'climate attribution' fit to inform legal and societal decisions will require enormous research effort."***

It is more difficult to make the case for 'usefulness'. None of the industry and government experts at the workshop could think of any concrete example in which an attribution might inform business or political decision-making. Especially in poor countries, the losses arising from extreme weather have often as much to do with poverty, poor health and government corruption as with a change in climate. The United Nations is planning to set up a fund with the aim of reducing loss and damage due to climate change, but the complexity of such issues is making negotiations difficult.

These caveats do not mean that event attribution is a lost cause. But they are a reminder that designers of climate services must think very clearly about how others might want to use the knowledge that climate scientists produce. That could be a task for social scientists, who have good methods for analysing decision-making and social transactions. They need to be more involved in shaping the production and dissemination of climate knowledge. ■

Return to sender

The bid to halt air transport of lab animals poses an imminent threat to biomedical research.

This week, the campaign group People for the Ethical Treatment of Animals (PETA) will take another step forward in its long-running, and increasingly successful, campaign to halt the transport by air of animals destined for the laboratory. It will announce that FedEx and UPS, the world's two largest cargo carriers, have written to it to affirm existing policies restricting the transport of most lab animals (see page 344). On the face of it, this seems pretty inconsequential. After all, neither carrier moves many research animals, and there are plenty of cargo firms that could make up any shortfall caused by PETA's pressure.

But appearances are deceptive: there could yet be an immediate and highly problematic effect. UPS has also said that it plans to change its policy soon to restrict the transport of amphibians, insects, crustaceans, molluscs and fish — all of which it allows at present. This could disrupt everything from the availability of the important frog model, *Xenopus* — three of whose major US-based suppliers rely on UPS next-day delivery — to the provision of the fruitfly *Drosophila* to international clients by the Bloomington Drosophila Stock Center at Indiana University.

And with PETA increasing the pressure, who is to say whether FedEx would not follow its arch-rival's lead and halt the transport of insects and other lower species? As with UPS, the effect would be huge. To name just a couple: FedEx currently ships fruitflies from suppliers including the Drosophila Species Stock Center at the University of California, San Diego, and Carolina Biological Supply in Burlington, North Carolina. The latter uses FedEx to ship *Drosophila*, along with crayfish, mussels and many other non-mammals, to science teachers.

If this is not enough to make scientists sit up and take notice, they might consider the use of lab rodents, now under threat in India from a PETA campaign to halt the transport of all research animals by Air India. The National Institute of Nutrition in Hyderabad, a major government supplier of specialized mice, relies on the airline. As PETA undertakes a systematic push to target all major cargo carriers, scientists in any country who rely on air freight to deliver rodents should be on notice that their turn may be next. Of course, in the increasingly global world of science it is already, in many senses, everyone's turn.

The pronouncements by FedEx and UPS, together with similar bans on animal movement made previously by airlines and ferry companies, are especially worrying because they indicate that biomedical researchers in many different countries, through reticence and passivity, are losing the battle for the hearts and minds of the public when it comes to the need for, and legitimacy of, animal research. Why else would high-profile companies be willing to indicate, however implicitly, that they want no part in a transportation infrastructure that is crucial to global biomedical science?

If individual scientists wait until they are personally affected — until the day when that mouse carefully bred in Shanghai or Singapore or Stockholm cannot be had for love nor money in San Francisco — it will be long past too late to mount the vigorous, public campaign in defence of animal research that is so sorely called for at this moment.

As researchers join this battle — and join it, they must — they should, as a first step, work through their institutions, academic societies and umbrella groups to make an urgent, articulate, unified case to UPS and FedEx that the shipping of animals, mammalian and otherwise, is essential for both biomedical research and scientific education. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

M. PANZERI

Social networks can spread the Olympic effect

The classic economic approach of using incentives is not always the best way to change human behaviour, argues **Paul Ormerod***.*

After a summer of sport, the London Olympics and Paralympics have ended and the city is now returning to normal. For London, normal means roads and public transport that are crammed, especially at peak times. It was all very different during the games, when many of the streets and shops in this dynamic city were eerily deserted. What made behaviours change so dramatically? And what lessons can be learned for behaviour change in other arenas?

Congestion was a potential major headache for the organizers of the Olympics. The conventional way to prompt a change in behaviour such as driving is to use incentives, the price mechanism beloved of economists. There is already a congestion charge for vehicles entering the city centre, so this could have been ramped up. And a special levy could have been introduced for travel on public transport.

But the increases in price would have had to be enormous to deter people, so London relied instead on social-network effects. Before the games, a massive publicity campaign focused on how crowded the centre of the city would be. Bus and train passengers, for example, were bombarded at regular intervals with recorded announcements from mayor Boris Johnson that warned just how busy public transport would be, and urged people to avoid them if they could.

The strategy worked — too well in fact — because of feedback effects. People do not receive such warnings as isolated individuals: they discuss them widely with friends and work colleagues. Employers reinforced the effect by promoting special arrangements for home-working and flexible hours. As a result, commuting cyclists had many roads to themselves and visitor numbers at flagship London venues fell by one-third.

TACKLING SOCIAL, ECONOMIC AND GLOBAL ISSUES WILL REQUIRE REAL, FUNDAMENTAL CHANGES TO BEHAVIOUR.

Johnson gave us a glimpse of public policy as it could be applied in the twenty-first century, relying on network effects rather than on incentives. In the twentieth century, both social and economic policy in the West were dominated by the principles of conventional economic theory: individuals with fixed tastes and preferences took decisions in isolation, and reacted to changes in incentives. So to achieve a policy goal, politicians would change tax rates and offer subsidies. This model is not wrong. But it is incomplete picture of the way in which the world now operates.

Network effects are not new. Throughout history, a crucial feature of human behaviour has been our propensity to copy or imitate the behaviours, choices and opinions of others. We can see it in the fashions in pottery in the Middle Eastern Hittite Empire of three-and-a-half millennia ago. But we are now much more aware of what other people are doing, or plan to do. For the first time in human history, more than half of us live in cities, in close, everyday proximity to large numbers of other people. And the Internet has revolutionized communication.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/imbtmq

Social networks are often thought of as a web-based phenomenon: Facebook, for example. Such forums can indeed influence behaviour. But real-life social networks — family, friends, colleagues — are even more important in helping to shape preferences and beliefs.

Social problems such as obesity are driven by network effects. It is not that people decide to copy fat friends and eat huge amounts; here, the network effect is one of peer acceptance. If most of your friends are obese, then it is more acceptable for you to put on weight. The problem of worklessness is also driven by networks. My home town of Rochdale, UK, attracted notoriety a couple of years ago when 84% of working-age adults on one council estate were found to be on benefits. Yet estates with very similar socioeconomic backgrounds had much lower rates, although still high by national standards. The social values of some estates had evolved to make being on benefits the norm.

A great deal of Europe's economic policy can be seen as an attempt by various players to use the social-network effect to get their narrative version of events to 'go viral' and dominate financial markets, almost without regard to objective reality. For example, although the United Kingdom has a higher public-sector debt relative to the size of its economy than, say, Spain, the United Kingdom is perceived as sound and Spain as risky.

Thanks to advances in network theory, we now know much more about how behaviour is spread and contained across networks than we did even ten years ago. Something that is particularly disturbing for policy-makers is the inherent level of uncertainty: some network effects simply fail to spread, and it is impossible to predict accurately how much traction an idea will get, and how any one event will unfold.

Tackling social, economic and global issues, such as climate change, will require real, fundamental changes to behaviour. To make this a reality, policy-makers, in both the public and the corporate sphere, will need to radically change their view of how the world operates. The inherent uncertainties of social networks make policies much harder to implement, so network theory must come up with effective, practical tools that help policy-makers to achieve their goals. For when they work, as we saw in London, social networks are a powerful and useful way to get things done. ■

Paul Ormerod *is an economist and complex-systems theorist with Volterra Consulting in London, and author of* Positive Linking: How Networks Can Revolutionise the World.
e-mail: pormerod@volterra.co.uk

RESEARCH HIGHLIGHTS

Selections from the scientific literature

ASTROPHYSICS

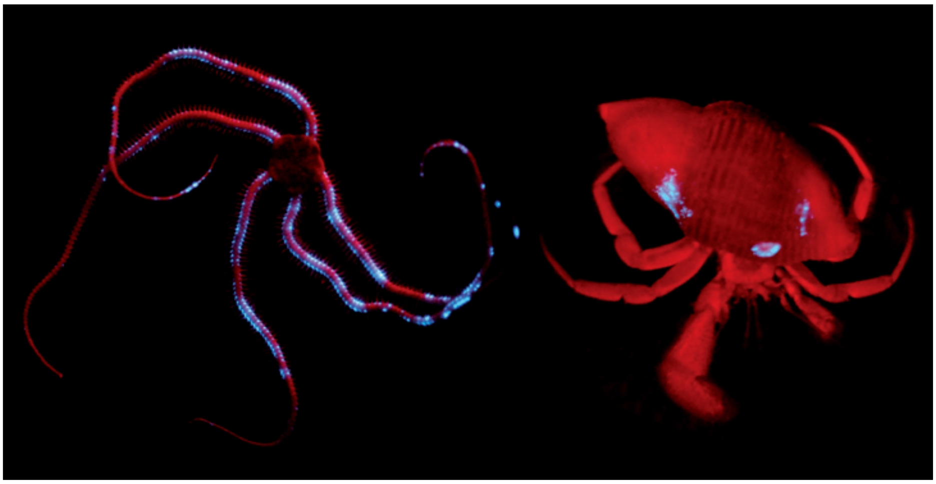
Stellar duo tests Einstein's theory

By studying the shrinking orbit of a pair of recently discovered white dwarf stars, astronomers have found further evidence that Einstein's theory of general relativity is correct.

The theory predicts that massive, accelerating objects like the two closely orbiting white dwarfs should emit gravitational waves — ripples in space-time that have never been detected directly. This release of energy, in turn, would cause the dwarfs' orbit to decay at a rate of around 0.26 milliseconds a year. James Hermes of the University of Texas at Austin and his colleagues used four telescopes to observe the dwarfs over 13 months. Their observations confirm that this is indeed roughly the rate at which the dwarfs are moving closer together.

Additional data would be needed to detect an orbital decay that deviates significantly from the rate predicted by general relativity, the team adds.

Astrophys J. 757, **L21 (2012)**



S. JOHNSEN

MARINE BIOLOGY

Glowing is rare on the sea floor

A survey of animals that live on the sea floor suggests that they are less likely to be bioluminescent than are species that swim freely at similar depths.

Sönke Johnsen at Duke University in Durham, North Carolina, and his colleagues dredged species (examples pictured) from 500–1,000 metres below the sea surface around the Bahamas, and examined them in tanks. Fewer than 20% of the creatures glowed, whereas roughly 80% of free-swimming species from similar depths are bioluminescent.

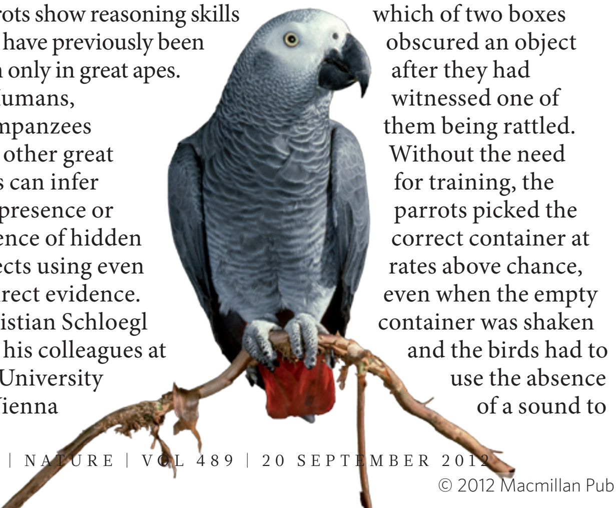
In a separate study, Johnsen, along with Tamara Frank of Nova Southeastern University in Dania Beach, Florida, and another colleague eased eight species of floor-dwelling crustacean up from the sea floor 500–700 metres down around the Bahamas and in the Gulf of Mexico.

The researchers kept the animals in light-tight boxes to protect their sensitive photoreceptors, and found that the creatures could best detect blue light — that is, wavelengths similar to those that filter through the water from the surface and are emitted by bioluminescent species. Some of the crustaceans had eyes that were sensitive to dim light, but were much less responsive to movement.

J. Exp. Biol. 215, **3335–3343; 3344–3353 (2012)**

ANIMAL COGNITION

Parrots can make inferences



S. DALTON/NATUREPL.COM

Parrots show reasoning skills that have previously been seen only in great apes.

Humans, chimpanzees and other great apes can infer the presence or absence of hidden objects using even indirect evidence. Christian Schloegl and his colleagues at the University of Vienna tasked six African grey parrots (*Psittacus erithacus*; **pictured**) with determining which of two boxes obscured an object after they had witnessed one of them being rattled. Without the need for training, the parrots picked the correct container at rates above chance, even when the empty container was shaken and the birds had to use the absence of a sound to guide their decisions.

Paying attention to sounds may be more important to parrots than to other animals that have failed the same test, including monkeys and dogs, the researchers suggest.

Proc. R. Soc. B 279, **4135–4142 (2012)**

NEUROSCIENCE

Memory boost with sleep

Using external stimulation to 'replay' recent experiences during sleep can strengthen the memories of those events, according to a study in rats.

Daniel Bendor and Matthew Wilson at the Massachusetts Institute of Technology in Cambridge trained rats to run to either the left or the right in response to one of two sounds, while recording from the brain's hippocampus. As the animals slept, the researchers played the sounds again to see whether this would trigger the rats to recall the task. The duo observed signs of 'replay' by analysing the response of neurons in the hippocampus, the brain region in which memories are thought to be consolidated.

These findings echo recent experiments in which human sensory learning improved following exposure to task-related cues during sleep, suggesting that such cues might strengthen memories.

Nature Neurosci. **http://dx.doi.org/10.1038/nn.3203 (2012)**

NANOTECHNOLOGY

Bigger rings allow thinner nanotubes

An ultrathin carbon nanotube that is stable up to temperatures of 1,000 kelvin has been predicted by a team at the Autonomous University of Madrid. The structure resembles a double helix, with alternating single, double and triple carbon–carbon bonds.

Eduardo Menéndez-Proupin and his colleagues simulated the structure and described the spectral signatures that would enable scientists to identify it experimentally. The predicted molecule is just 0.32 nanometres in diameter. Standard nanotubes this thin have never been observed, because the carbon bonds in their six-atom rings have large distortions and become destabilized as the nanotubes get thinner. The carbon rings in the predicted molecule are larger than in standard nanotubes, and the bonds are arranged such that they are less strained.

Phys. Rev. Lett. 109, **105501 (2012)**

NEUROSCIENCE

Social isolation thins neural sheath

Mice raised in isolation are unsociable and slow to learn complex tasks as adults. Failure to develop normal coatings around neurons during a crucial growth period could explain these deficits.

Gabriel Corfas at Boston Children's Hospital in Massachusetts and his colleagues found that mice isolated between 21 and 35 days old were particularly vulnerable to lasting effects. In these mice, oligodendrocytes — cells that produce the fatty layers which sheath neurons to facilitate electrical signalling — made abnormally thin sheaths in the prefrontal cortex, an area linked to sociability and memory.

In the same brain area, isolated mice showed reduced levels of NRG1, a protein that has a role in oligodendrocyte development. Mice engineered to lack an NRG1 receptor in oligodendrocytes mimicked the negative effects of isolation, suggesting that social experience might affect neural development through the NRG1 signalling pathway.

Science 337, **1357–1360 (2012)**

PLANETARY SCIENCE

Volcanic signs in Martian clays

Clay minerals on the surface of Mars (**pictured**) could be signs of previous volcanic activity rather than an indication that the planet had a warm and wet climate in the past, as has been assumed.

Clays can form when igneous rock is altered by water present at the surface or underground. But Alain Meunier at the University of Poitiers in France and his colleagues suggest that the Martian clays could have precipitated directly out of a water-rich magma, which filled voids in the igneous rock as it cooled. When the researchers analysed rocks from terrestrial lavas from a French Polynesian atoll, they found similar spectral signatures to those of the Martian clays.

The authors' suggestion — soon to be investigated by the Mars rover Curiosity — is that the planet's early climate was volcanic, but not necessarily wet.

Nature Geosci. **http://dx.doi.org/10.1038/ngeo1572 (2012)**

COMMUNITY CHOICE

The most viewed papers in science

GEOLOGY

One million years of rubbing rocks

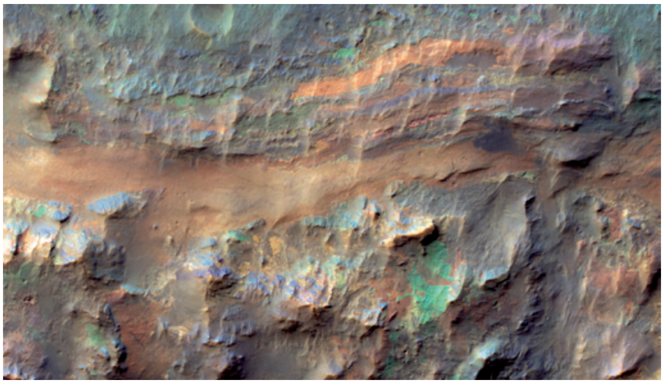
★ HIGHLY READ on geology.gsapubs.org in August

The strangely smooth shape of boulders in Chile's Atacama Desert, one of the world's driest locations, could be due to the rocks rubbing against each other during earthquakes.

Jay Quade at the University of Arizona in Tucson and his colleagues analysed the patterns of erosion shown by the boulders. The researchers determined that the rocks' smooth sides, and depressions in the sediment around them, could be best explained by rubbing and rocking motions experienced during earthquakes. In February 2010, two members of the team were present when an earthquake with a magnitude of 5.2 struck about 100 kilometres from their location, enabling them to observe the rocks rubbing against each other for about a minute.

Earthquakes of a similar or larger magnitude occur roughly once every four months, and the authors calculate that the boulders could have experienced 40,000–70,000 hours of rubbing over the past 1.3 million years.

Geology 40, **851–854 (2012)**



HIRISE/UNIV. ARIZONA/JPL/NASA

ORGANIC CHEMISTRY

Tagging molecules with fluorine

Attaching fluorine atoms to organic molecules is important in, for example, tweaking the properties of a drug candidate. John Groves at Princeton University in New Jersey and his colleagues have now discovered a way to substitute fluorine atoms at previously inaccessible positions in a molecule: carbon–hydrogen bonds, which are notoriously unreactive.

The researchers used a manganese porphyrin catalyst to assist the reaction, which they say requires only simple apparatus and mild conditions. They were able to fluorinate simple hydrocarbons and even complex steroid molecules with yields of up to 60%.

The technique could be used to incorporate radioactive fluorine into a wide range of biomolecules for imaging.

Science 337, **1322–1325 (2012)**

➲ NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

US budget cuts

US science agencies would have their funding frozen at 2012 levels until 27 March 2013, under a bill passed by the House of Representatives on 13 September. The 'continuing resolution' provides temporary funds after the end of the 2012 US fiscal year (30 September), as US politicians are yet to agree on a 2013 budget. But it ignores another looming deadline: an enforced, across-the-board budget cut (or 'sequester') in January 2013 that would reduce non-defence spending at science agencies by 8.2%, according to an analysis released on 14 September by the White House's Office of Management and Budget. Politicians will have scant time after November's federal election to work out how to dodge this cut. See go.nature.com/def7mf for more.

Nuclear wind down

Japan is to phase out its 50 remaining nuclear reactors by the 2030s, the country's government announced on 14 September. In contrast to Germany, which last year decided to phase out its 17 nuclear reactors by 2022, the Japanese phase-out period is relatively long and so leaves scope for shifts or reversals in the future. See go.nature.com/dx6x4n for more.

M. EMETSHU

New monkey species discovered

A slender, golden-maned monkey, well known to Congolese villagers in the Lomami River basin and first sighted by researchers in 2007, has been declared a new species. In an article published in *PLoS ONE* on 12 September (J. A. Hart *et al. PLoS ONE* **7,** e44271; 2012), Kate Detwiler of Florida Atlantic University in Boca Raton and her colleagues identified *Cercopithecus lomamiensis* from facial, behavioural and genetic features that distinguish it from similar species. It is only the second African monkey species discovered in 28 years.

Drug-safety drive

The European Union (EU) agreed to change its drug-oversight rules on 11 September, which should lead to speedier safety assessment and withdrawal of a drug from sale across the EU if its safety is questioned in one member state. The move was prompted by the scandal over the tardy withdrawal of the diabetes medicine Mediator (benfluorex) that erupted in France in November 2009. The drug, made by the French firm Servier, based in Neuilly-sur-Seine, was widely prescribed to people without diabetes as an appetite suppressant and is estimated to have caused more than 2,000 deaths from heart-valve disease and high blood pressure. See go.nature.com/9hukkh for more.

Open-access call

To mark the tenth anniversary of the Budapest Open Access Initiative — a campaign calling for open access to all peer-reviewed research — supporters on 12 September released a set of recommendations to cover the next ten years (see go.nature.com/kfvbbv). They called on institutions and funding agencies to introduce open-access policies, and they set a new goal: to make open access the default method for distributing peer-reviewed research everywhere within a decade.

Space-agency boost

Poland has joined the European Space Agency (ESA). At a ceremony in Warsaw on 13 September, the nation became the 20th country to join the agency. Poland has been cooperating with ESA since 1994, and is involved in several of its science missions, including the Rosetta spacecraft currently en route to the comet Churyumov–Gerasimenko. Poland will officially become a full member state at an ESA council meeting in November.

Nature reserve

The largest man-made coastal reserve in Europe is to be built using 4.5 million tonnes of earth from the construction of tunnels for London's Crossrail network. The project, launched on 17 September and managed by the Royal Society for the Protection of Birds, will transform 670 hectares of farmland on Wallasea Island in the Thames Estuary into wetlands and marshes to attract and house birds and other wildlife.

BUSINESS

Sequencing power

One of a new generation of DNA sequencers, a benchtop device from US biotech firm Life Technologies, began shipping to customers on 13 September, the company says. The US$150,000 'Ion Proton' uses $1,000 chips to sequence between 60 million and 80 million DNA fragments, each up to 200 bases long, in 4 hours. The company, based in Carlsbad, California, says that next year it will release a chip that can sequence a full human genome within 4 hours. Its competitor Illumina, headquartered in San Diego, California, says that its own high-throughput instrument, which can complete a full human genome within 24 hours, will be available by the end of this year. See go.nature.com/8g54pj for more.

Genomics deal

The world's largest genome sequencing centre, BGI in Shenzhen, China, announced plans on 17 September to buy into a leading human-genome sequencing firm. In a deal worth US$117.6 million, BGI plans to buy all the outstanding shares of Complete Genomics, which is based in Mountain View, California. Complete Genomics provides services for academic, medical and industry customers, but reported $19 million in losses in the second quarter of 2012 and in June laid off 55 employees.

PEOPLE

Murder plea

Amy Bishop, a biologist formerly at the University of Alabama, Huntsville, who shot and killed her department chairman and two other colleagues during a faculty meeting on 12 February 2010 (see *Nature* **465,** 150–155; 2010), has pleaded guilty to murder. The plea, made in Madison County circuit court in Edwardsville, Illinois, on 11 September, means that Bishop will probably spend the rest of her life in prison but avoid the death penalty. See go.nature.com/ldk3g1 for more.

Drug doyen

Neurogeneticist Christopher Austin will head the National Center for Advancing Translational Sciences (NCATS), Francis Collins, director of the National Institutes of Health (NIH), announced on 14 September. Austin (**pictured**) worked in drug discovery at Merck before moving to the NIH in 2002, and has been head of the NCATS division of preclinical innovation since the centre, based in Bethesda, Maryland, was launched in December 2011 with a proposed US$575-million budget. See go.nature.com/fmdker for more.

M. BARTLETT/NHGRI

RESEARCH

Himalayan melt

Glacier melt is accelerating in eastern and central parts of the Himalayas, but glaciers in the west could be growing, according to a report from the US National Research Council published on 12 September. Seasonal glacial melt contributes to the water supply of some 1.5 billion people in the region, but the report suggests that retreating glaciers are unlikely to threaten those supplies in the next few decades.

King question

A skeleton that may be that of Richard III, fifteenth-century king of England, has been discovered in Leicester, UK. Archaeologists from the University of Leicester last week announced that they had found the remains of a male with skull trauma and severe scoliosis — curvature of the lower back — consistent with historical accounts of the king's appearance and death. The body was found during a dig on the site of Grey Friars Church, now a city-centre car park. Researchers plan to compare mitochondrial DNA from the skeleton with that of Michael Ibsen, a man believed to be a descendant of the king's sister.

COMING UP

22–25 SEPTEMBER
The European Molecular Biology Organization meets in Nice, France, with keynote presentations on the life and death of RNA, and on the dynamic genome.
go.nature.com/rj4a12

26–28 SEPTEMBER
In Arusha, Tanzania, politicians, firms, farmers and scientists convene at the African Green Revolution Forum to discuss how to invest in the continent's agricultural development.
go.nature.com/x3ymat

Weather satellite

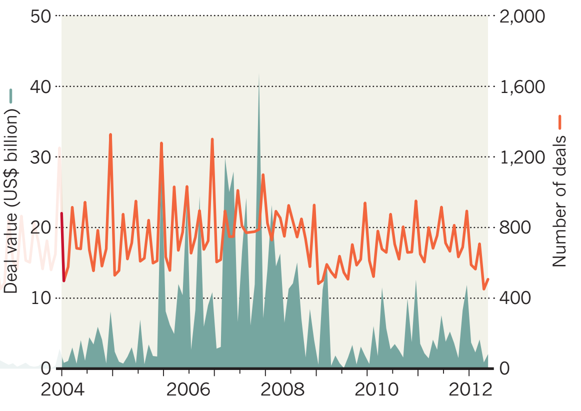
The meteorological satellite MetOp-B was launched from Baikonur in Kazakhstan by the European Space Agency (ESA) on 17 September. It is set to replace the MetOp-A satellite by the end of 2012. MetOp-B will measure ozone and other trace gases, but it lacks most of the climatic instruments of ESA's satellite Envisat — lost in April — and the agency's as-yet unlaunched Sentinel satellite system, which faces delays owing to funding negotiations (*Nature* **484,** 423–424; 2012).

TREND WATCH

The financial and economic crisis has caused a worldwide drop in spending on research and development, as well as reductions in the creation of companies and venture-capital investment (see chart). Figures in the Organisation for Economic Co-operation and Development's 2012 outlook report, published on 13 September, also suggest that China, South Korea and other emerging Asian economies are out-innovating the Western world. See go.nature.com/p1quzi for more.

DEAL DOWNTURN

The global economic crisis has cut into venture-capital investments over the past four years.



SOURCE: OECD/THOMSON ONE

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

RESEARCH ANIMALS PETA campaign closes off more shipping options **p.344**

ETHICS Anaesthesiology wakes up to record-setting fraud **p.346**

EPIDEMIOLOGY Does West Nile disease damage the kidneys? **p.349**

NEUROSCIENCE The remarkably busy lives of idle brains **p.356**

At 500,000 years, the dating of this skull of *Homo heidelbergensis* clashed with previous DNA dates for Neanderthal origins.

J. TRUEBA/MSF/SPL

ANTHROPOLOGY

Studies slow the human DNA clock

Revised estimates of mutation rates bring genetic accounts of human prehistory into line with archaeological data.

BY EWEN CALLAWAY

The story of human ancestors used to be writ only in bones and tools, but since the 1960s DNA has given its own version of events. Some results were revelatory, such as when DNA studies showed that all modern humans descended from ancestors who lived in Africa more than 100,000 years ago. Others were baffling, suggesting that key events in human evolution happened at times that flatly contradicted the archaeology.

Now archaeologists and geneticists are beginning to tell the same story, thanks to improved estimates of DNA's mutation rate — the molecular clock that underpins genetic dating[1–4]. "It's incredibly vindicating to finally have some reconciliation between genetics and archaeology," says Jeff Rose, an archaeologist at the University of Birmingham, UK. Archaeologists and geneticists may now be able to tackle nuanced questions about human history with greater confidence in one another's data. "They do have to agree," says Aylwyn Scally, an evolutionary genomicist at the Wellcome Trust Sanger Institute in Hinxton, UK. "There was a real story."

The concept of a DNA clock is simple: the number of DNA letter differences between the sequences of two species indicates how much time has elapsed since their last common ancestor was alive. But for estimates to be correct, geneticists need one crucial piece of information: the pace at which DNA letters change.

Geneticists have previously estimated mutation rates by comparing the human genome with the sequences of other primates. On the basis of species-divergence dates gleaned — ironically — from fossil evidence, they concluded that in human DNA, each letter mutates once every billion years. "It's a suspiciously round number," says Linda Vigilant, a molecular anthropologist at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. The suspicion turned out to be justified.

In the past few years, geneticists have been able to watch the molecular clock in action, by sequencing whole genomes from dozens of families[5] and comparing mutations in parents and children. These studies show that the clock ticks at perhaps half the rate of previous estimates, says Scally.

In a review published on 11 September[1], Scally and his colleague Richard Durbin used the slower rates to reevaluate the timing of key splits in human evolution. "If the mutation rate is halved, then all the dates you estimate double," says Scally. "That seems like quite a radical change." Yet the latest molecular dates mesh much better with key archaeological dates.

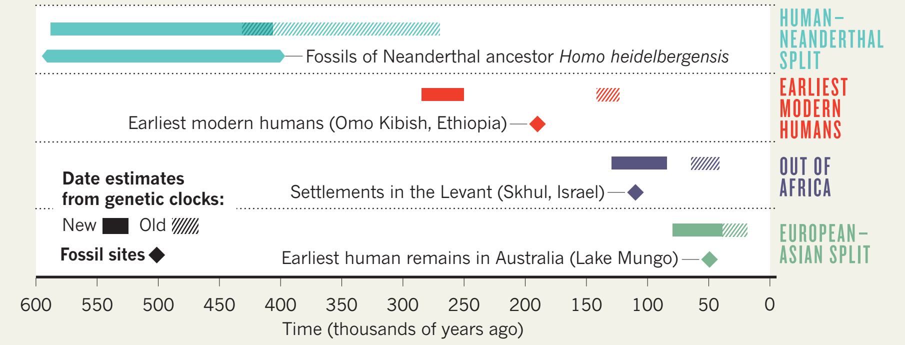
Take the 400,000–600,000-year-old Sima de Los Huesos site in Atapuerca, Spain, which yielded bones attributed to *Homo heidelbergensis*, the direct ancestors of Neanderthals. Genetic studies have suggested that earlier ancestors of Neanderthals split from the branch leading to modern humans much more recently, just 270,000–435,000 years ago. A slowed molecular clock pushes this back to a more comfortable 600,000 years ago (see 'Better agreement over the human story').

A slower molecular ▶

➲ **NATURE.COM**
For a special issue on ancient human migrations, visit:
go.nature.com/jcunxr

BETTER AGREEMENT OVER THE HUMAN STORY

Dates estimated from DNA evidence conflicted with those from fossil sites that document key events in prehistory, but dates gained using a slower DNA clock are resolving some conflicts.



SOURCE: REF. 1

▶ clock could also force scientists to re-think the timing of later turning points in prehistory, including the migration of modern humans out of Africa. Genetic studies of humans around the world have suggested that the ancestors of Europeans and Asians left Africa about 60,000 years ago. That date caused many to conclude that 100,000-year-old human fossils discovered in Israel represented a dead-end migration rather than the beginning of a global exodus, says Scally. Scally's calculations put "out of Africa" closer to 120,000 years ago, suggesting that the Israeli sites represent a launching pad for the spread of humans into Asia and Europe.

The latest genetic dates also fit with several sites in the Middle East that contain tools apparently made by modern humans but dating to around 100,000 years ago. At that time, sea levels between Africa and the Arabian Peninsula were lower than they are now, and a wetter climate would have made the peninsula lush and habitable, perhaps beckoning modern humans out of Africa. Rose, who works one such site, in Oman, says that he "has been over the moon" since reading Scally and Durbin's paper.

The revised molecular clock may also help to settle a debate over whether humans ventured further into Asia more than 60,000 years ago, says Michael Petraglia, an archaeologist at the University of Oxford, UK, who favours an early date.

Although a slowed molecular clock may harmonize the story of human evolution, it does strange things when applied further back in time, says David Reich, an evolutionary geneticist at Harvard Medical School in Boston, Massachusetts. "You can't have it both ways."

For instance, the slowest proposed mutation rate puts the common ancestor of humans and orang-utans at 40 million years ago, he says: more than 20 million years before dates derived from abundant fossil evidence. This very slow clock has the common ancestor of monkeys and humans co-existing with the last dinosaurs. "It gets very complicated," deadpans Reich.

Some researchers, including Scally, have proposed that the mutation rate may have slowed over the past 15 million years, thereby accounting for such discrepancies. Fossil evidence suggests that ancestral apes were smaller than living ones, and small animals tend to reproduce more quickly, speeding the mutation rate.

Little concrete evidence supports this idea, says Reich. He agrees that the molecular clock must be slower than was thought, but says that the question is how slow. "My strong view right now is that the true value of the human mutation rate is an open question." ■

1. Scally, A. & Durbin, R. *Nature Rev. Genet.* **13,** 745–753 (2012).
2. Langergraber, K. E. *et al. Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1211740109 (2012).
3. Hawks, J. *Proc. Natl Acad. Sci. USA* http://dx.doi.org/10.1073/pnas.1212718109 (2012).
4. Sun, J. X. *et al. Nature Genet.* http://dx.doi.org/10.1038/ng.2398 (2012).
5. Kong, A. *et al. Nature* **488,** 471–475 (2012).

RESEARCH ANIMALS

Lab-animal flights squeezed

Two biggest cargo carriers affirm that they will not ship mammals and non-human primates, as activist pressure mounts to stop research-animal airlifts.

BY MEREDITH WADMAN

For researchers who rely on lab animals shipped from distant sources, and for the companies that breed them, the options are narrowing again. This week, People for the Ethical Treatment of Animals (PETA) will announce that it has obtained written assurances from the world's two largest air-cargo carriers, FedEx and UPS, that they will not transport mammals for laboratory use. UPS says that it is also planning to further "restrict" an exemption that allows the transport of amphibians, fish, insects and other non-mammals.

Neither company currently ships large numbers of lab animals. But PETA, an activist group based in Norfolk, Virginia, sought the carriers' written assurances as a way to foreclose alternatives for lab-animal breeders and their customers, who are increasingly being confronted with bans on transport by passenger airlines. "FedEx and UPS were not transporting many or any animals, but we felt it was crucial to go to them and discuss this as we knew that facilities trying to send non-human primates and other species would be going to them soon, as more and more passenger airlines refused to do business with them," says Kathy Guillermo, PETA's senior vice-president for laboratory investigations.

The commitments will have a direct impact on some researchers. "I am deeply concerned," says Darcy Kelley, a neurobiologist at Columbia University in New York City, who studies neural and muscular systems involved in vocal communication in the frog *Xenopus*. The supply companies that Kelley uses — Nasco in Fort Atkinson, Wisconsin; Xenopus One in Dexter, Michigan; and Xenopus Express of Brooksville, Florida — all ship the amphibians by air with UPS for next-day delivery. Losing access to the frogs because of shipping hurdles "would set my research back years", says Kelley. "It takes *Xenopus* females two years to get to sexual maturity. And maintaining an animal colony is a very expensive proposition."

For those who study mammals, the FedEx and UPS policies may have little immediate impact. The two companies are not used to ship non-human primates internationally, says

Michael Hsu, president of Shared Enterprises in Richlandtown, Pennsylvania, which maintains a macaque-breeding colony in Shanghai and imports research animals to the United States by air. In the United States, many other lab animals are domestically bred and shipped by truck. But although the FedEx and UPS declarations may be largely symbolic, they suggest that research advocates are failing to make the case for the use of lab animals, and they mark another success for groups such as PETA.

Many large passenger carriers will no longer transport non-human primates after being confronted by PETA and other animal activist groups (see *Nature* **483,** 381–382; 2012). United Airlines and Air France are among the few that have not ruled out primate transport. Air Canada is petitioning the Canadian Transportation Agency for permission to stop the practice. Now, PETA is extending its campaign to other species and to cargo carriers. Non-air transport across international borders is also under pressure. In March, the last two ferry companies transporting laboratory rodents into the United Kingdom said that they were stopping the practice.

FedEx, based in Memphis, Tennessee, says that its commitment not to ship animals reflects a policy that is at least five years old. "There was an active decision made that, especially here in the United States, that's just not how we wanted to do business," says Shea Leordeanu, manager of global public relations for the company. FedEx, the leading global cargo shipper, does occasionally transport animals — for example, it delivered horses to the equestrian events at the London Olympics — but only with special dispensation. Under such exemptions in recent years, as many as several dozen international shipments of research mice have travelled by FedEx annually, Leordeanu says. "However, FedEx has not transported any mice at all in many months," she adds, because customers have not requested its services.

UPS, based in Atlanta, Georgia, has limited animal shipments for more than a decade. With rare exceptions, it ships only amphibians, crustaceans, fish, insects, molluscs and certain lizards and turtles. "We currently are in the process of putting procedures in place to restrict those shipments as well," says Norman Black, director of global media services for UPS, but "the fact that we're considering restrictions doesn't mean a flat ban". The company's policy, he says, is "based both on our sustainability principles and on our marketing decisions. We do not consider animal shipments to be a target market for us, either economically or operationally."

Losing the option of shipping frogs by UPS would be "huge" for his company, says Burley Lilley, president of Xenopus Express, which serves around 100 academic customers throughout the United States. "Part of the reason our business is so good and the animals get there alive is because we use UPS."

PETA

Research beagles being air-freighted by Lufthansa before the carrier changed its policy.

Charles Hewett, executive vice-president and chief operating officer at the Jackson Laboratory in Bar Harbor, Maine, says that less than 10% of the several million specialized mice that Jackson ships from its US locations each year travel by air; most are shipped domestically by 18-wheel truck. The laboratory also breeds highly requested strains of mice at facilities overseas, so that they can be delivered quickly by truck.

"We do not use FedEx, we do not use UPS and in fact we believe very strongly that our mice should only be handled by truckers who have been trained to understand the animals' requirements," says Hewett.

Nonetheless, Hewett says he finds it "troubling that the corporate leaderships of UPS, FedEx and others yield to the pressure of a small minority who overlook the importance of what we do for preventing, curing and treating human disease."

"We do not consider animal shipments to be a target market for us, either economically or operationally."

For many of its international shipments, Jackson uses a contractor, Charles River Laboratories in Wilmington, Massachusetts, which did not respond to requests for comment. The PETA campaign has had an impact on Charles River in at least one instance. In 2010, less than 24 hours after PETA published a photo of beagles in the cargo hold of a Lufthansa airliner at New York's JFK airport, the German airline said that it would no longer ship dogs and cats for research. The dogs were in transit from research-animal breeder Marshall BioResources in North Rose, New York, to a Charles River Laboratories facility in Scotland.

PETA says that it is systematically approaching every major cargo carrier in the world, putting pressure on both international and domestic shipments. In India, for example, the government's National Institute of Nutrition (NIN), in Hyderabad, relies on Air India to ship specialized mouse strains to researchers and companies throughout the country. "From Hyderabad to Delhi by train would take more than 30 hours" and require an attendant, says Madan Chaturvedi, dean of life-sciences research at the University of Delhi. Without Air India transporting the animals, research at his institution "would definitely suffer", he says.

In response to pressure from PETA-India, Air India wrote to the group in July saying "we ... do not accept animals for experimental purposes." On 23 August, Air India issued a circular to all its managers and cargo staff declaring "Air India does not carry 'Live Animals for experimental purposes'". But Kalpagam Polasa, acting director at the NIN, told *Nature* last week that weekly flights of her institute's animals on Air India continue, labelled in the 'live animal' category, and costing her institute three times as much as previously. The airline did not respond to requests for comment.

Many scientists may shrug their shoulders at the personal impact of the trend in cargo-carrier policies, says Joseph Haywood, vice-president for science policy at the Federation of American Societies for Experimental Biology in Bethesda, Maryland, and vice-president for regulatory affairs at Michigan State University in East Lansing, where he is responsible for animal transport for the university. But, he says, "when they need that specific animal model to ask a critical question, they need to have that model. It could be across the street or across the world. We are moving to global science." ■ SEE EDITORIAL P.366

Breathe deeply: anaesthetists are well placed to conduct clinical trials with relatively little oversight.

T. PANNELL/CORBIS

MISCONDUCT

Retraction record rocks community

Anaesthesiology tries to move on after fraud investigations.

BY DAVID CYRANOSKI

One of the biggest purges of the scientific literature in history is finally getting under way. After more than a decade of suspicion about the work of anaesthesiologist Yoshitaka Fujii, formerly of Toho University in Tokyo, investigations by journals and universities have concluded that he fabricated data on an epic scale. At least half of the roughly 200 papers he authored on responses to drugs after surgery are in line for retraction in the coming months.

Like many cases of fraud, this one has raised questions about how the misconduct went undetected for so long. But the scope and duration of Fujii's deception have shaken multiple journals and the entire field of anaesthesiology, which has seen other high-profile frauds in the past few years.

Fujii, who could not be contacted for this article, was dismissed from Toho University in February because he lacked proper ethics approval for clinical studies that were detailed in eight papers. But suspicions about his entire 20-year publication record had been growing since 2000, when Peter Kranke, an anaesthesiologist at University Hospital Würzburg in Germany, first started to question Fujii's superhuman publication rate.

In some years, Fujii published more than a dozen randomized clinical trials that purported to test the efficacy and side effects of drugs such as granisetron, given to reduce nausea and vomiting after surgery. "It's impossible to publish so many," says Kranke. "If you just look at mere output, everybody who has performed at least one clinical trial should have some suspicion."

Fujii's data were also "too perfect", he says. Kranke analysed 47 of Fujii's articles on granisetron, published between 1994 and 1999, and found that the frequency of headaches — a common side effect of the drug — was identical or nearly identical in a suspiciously high number of groups involved in the trials[1].

At the time, Fujii responded merely by saying that he stood by his data[2], which seemed to show that granisetron had fewer side effects than other anti-emetic drugs. "We were disappointed that the journal accepted that," says Kranke. "Editors and peer reviewers advised us to pursue more worthwhile endeavours, rather than whistle-blowing. But it wasn't just whistle-blowing — we wanted people to know that" granisetron wasn't necessarily better than alternatives.

In the following years, similar doubts emerged about Fujii's studies of other drugs, such as the anti-emetic ramosetron. Yet the deception began to unravel only in September 2011, after growing doubts among anaesthesiologists prompted Toho University to begin an investigation into Fujii's lack of approval from institutional review boards for several clinical trials.

➲ NATURE.COM
Read more about a surge in retractions:
go.nature.com/cy8atp

Fujii's subsequent dismissal was soon followed by a flood of damning evidence about his work. On 8 March, *Anaesthesia* published an analysis[3] by John Carlisle, a consultant anaesthetist at Torbay Hospital in Torquay, UK, finding that 168 of Fujii's papers had results with "likelihoods that are infinitesimally small". One month later, 23 anaesthesiology journal editors wrote to the heads of six universities and medical centres with which Fujii had been affiliated, notifying them that 193 papers would be retracted unless the institutions could vouch for the data.

Five of those institutions have responded to say that they could not find evidence to corroborate the veracity of 88 papers. The sixth institution, the University of Tsukuba, has so far found only five papers to be valid. It is still investigating another 92 publications.

Late last month, Steven Shafer — editor-in-chief of *Anesthesia & Analgesia* — who has led the campaign to examine Fujii's papers, posted online the responses from the institutions, along with a notice of retraction for three papers in his own journal. Shafer expects other editors to begin retracting the remaining fraudulent papers, probably including those being considered by Tsukuba, in the forthcoming issues of their journals.

Meanwhile, an investigation begun in March by the Japanese Society of Anesthesiologists reported in June that 172 of Fujii's papers were probably fraudulent, in some cases because there was no evidence that the data had actually been collected. On 30 August, the society announced that Fujii, who had already left the society of his own volition, would be permanently barred. All of Fujii's co-authors have denied knowledge of his wrongdoing, according to Koji Sumikawa, a member of the society's board of directors who was involved in the investigation, and an anaesthesiologist at Japan's Nagasaki University School of Medicine.

Sumikawa says that Fujii's work has had little impact on clinical practice, because the anti-emetics he studied are rarely used in Japan. But it has embarrassed the field of anaesthesiology, which was already reeling from two high-profile fraud cases. In 2009, 21 publications by Scott Reuben, who was based at Tufts University School of Medicine in Boston, Massachusetts, were retracted because they contained fabricated data[4]. The following year, around 90 papers by Joachim Boldt, formerly of the Ludwigshafen Hospital in Germany, were retracted from 11 journals, because of fabrication and because Boldt did not have proper ethics approval for the trials.

The cases have prompted a spate of articles in anaesthesiology journals lamenting the scale of the frauds, and discussing ways to avoid similar incidents. In an letter sent to journal subscribers in March, Shafer said that he regretted that the journal had not investigated further after Kranke's original article (Shafer was not the editor-in-chief at that time). "The journal's response to the allegations of research fraud … was inadequate," he wrote. "The subsequent submissions to the Journal by Dr. Fujii should not have been published without first vetting the allegations of fraud."

Kranke thinks that anaesthesiologists are now more attuned to the possibility of misconduct, and that journal editors are much more willing to act on allegations. "After the Boldt and Reuben cases, it became fashionable to dig up these things."

AMPLE OPPORTUNITIES

Most anaesthesiologists insist that there is no evidence that their field is more prone to fraud than any other. But Carlisle says that anaesthesiology does offer many opportunities to generate large sets of clinical data very quickly. Millions of anaesthetic procedures are performed every year during surgeries, and patient outcomes are immediate and easy to measure. There are "frequent opportunities for anaesthetists to conduct clinical studies very quickly, potentially by themselves, without overview from other people", he says. "This might contribute to greater opportunities for them to succumb to the temptation of fraud."

How did Fujii get away with his deception for so long? One reason could be that he spread his publications over a wide range of journals, in fields as diverse as gynaecology and ophthalmology, suggests Sumikawa. "No one is looking at all of these," he says.

Fujii's peripatetic career may have also provided a smokescreen for his fraudulent behaviour. Over two decades, he held posts at five institutions, and adjunct positions at two more, making it easy for him to claim that data had been generated or ethics approval had been granted while he was in a previous post. If any of Fujii's colleagues were suspicious, they did not come forward at the time, says Sumikawa, who plans to set up a mechanism for whistle-blowers to report concerns about colleagues.

Kranke says he is pleased that his field is finally focusing its attention on misconduct. He is convinced that although Fujii's is an exceptional case, the researcher cannot be written off as merely a "bad apple". "It's a system failure," he says. ■ SEE EDITORIAL P. 335

1. Kranke, P., Apfel, C. C. & Roewer, N. *Anesth. Analg.* **90,** 1004–1007 (2000).
2. Fujii, Y. *Anesth. Analg.* **90,** 1004–1007 (2000).
3. Carlisle, J. B. *Anaesthesia* **67,** 521–537 (2012).
4. White, P. F., Roscow, C. E. & Shafer, S. L. *Anesth. Analg.* **112,** 512–515 (2011).

INDUSTRY

UK technology-boost plan disappoints

Government strategy to support industries of the future has little cash to back the vision.

BY ANANYO BHATTACHARYA

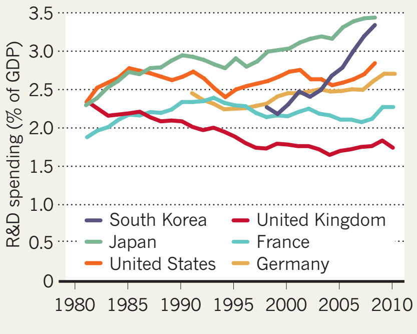
Industrial policy has long been anathema in UK politics. Anyone proposing a comprehensive strategy to spur the growth of high-tech industries would be accused of 'picking winners' — a euphemism for shoring up failing businesses with subsidies. But in the face of Britain's sagging economy, some of its leaders may have had a change of heart.

In a speech on 11 September, UK business secretary Vince Cable set out a long-term vision for how the government could boost established powerhouses of the British economy, such as the car and aerospace industries, and emerging areas, such as biotechnology. But although they welcome his words, science-policy experts say that the measures announced are too modest and piecemeal to succeed given the meagre government spending on applied research. By contrast, federal and state governments in the United States liberally support technology start-ups, and huge, mission-driven agencies such as NASA have spurred the growth of technology-based companies.

Others say that flat public funding for UK research is the real barrier for science-driven businesses. Geneticist Paul Nurse, president of the Royal Society, says that the minister's vision failed to acknowledge that a successful industrial policy rests on long-term, stable support for science. "The discovery science lays the foundation of the building, and if you don't have the foundation it will eventually fall over."

LEFT BEHIND

UK spending on research and development (R&D) has fallen as a proportion of gross domestic product (GDP) because of sagging business investment.



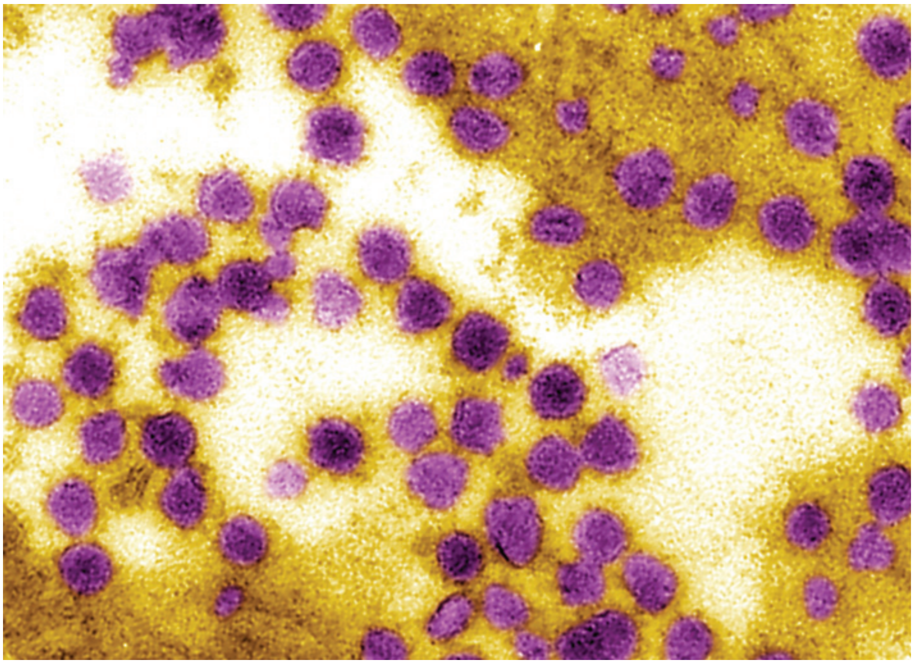
SOURCE: EUROSTAT

Cable's plans include a government-backed bank that would lend to small businesses struggling to raise cash, and a £250-million (US$405-million) 'employer ownership' pilot scheme to help companies build a skilled workforce through apprenticeships and vocational courses. Competitions for £1.25 million in funds for energy-efficient computing and £1 million for technologies to extend the life of batteries by harvesting energy from the environment will be launched on 8 October. Open to universities and businesses, they will award cash to study how innovations might be commercialized. Cable also announced a new Innovation and Knowledge Centre in Synthetic Biology. Details are hazy, but it is likely to be somewhere that researchers and entrepreneurs can meet to work out ways of overcoming roadblocks to commercializing science.

Critics say that these measures aren't enough to halt the fall in spending on research and development (see 'Left behind'), especially by UK businesses. "Given the total inadequacy of the resources being made available, it's difficult to see this strategy making any difference," says Richard Jones, pro-vice chancellor for research and innovation at the University of Sheffield. He points out that his university's Advanced Manufacturing Research Centre, praised for its collaborative research with engineering firms in a paper accompanying Cable's speech, would not have got off the ground without public funds from the European Union and elsewhere.

Many policy experts advocate boosting the budget of the Technology Strategy Board — an agency that supports near-market research and development — and of the seven 'Catapult' centres, loosely modelled on Germany's Fraunhofer Institutes, which aim to stimulate links between businesses and universities. But with Britain in deep recession, any major investment in industrial research would face opposition from politicians on the right, who often argue that the best thing a government can do for business is to spend less and cut red tape.

"There are absolutely areas where we believe there should be deregulation and simplification," says Steve Bates, chief executive of the UK BioIndustry Association in London. "On the other hand, you don't get strategic sectors growing overnight without support and nurturing. The two go together." ■



SGO/BSIP/CORBIS

The fuzzy bodies of West Nile virus, which may target the kidneys.

EPIDEMIOLOGY

The hidden threat of West Nile virus

Researchers probe possible link with kidney disease.

BY AMY MAXMEN

This year is on track to be the worst on record for West Nile virus in the United States. As of 11 September, more than 2,600 new cases, including 118 deaths, had been reported from across the country to the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia.

Symptoms of the mosquito-borne disease range from none (in most people) to life-threatening brain inflammation, and it can leave survivors with long-term disabilities including paralysis and fatigue. Researchers are now investigating suggestions that even mild infections may leave another lasting burden — kidney disease.

“We are early in our understanding, but this really worries me,” says Kristy Murray, an epidemiologist and clinical researcher at Baylor College of Medicine in Houston, Texas, who has found hints that the virus may persist in the kidney long after the initial infection. This week she is moving her work on the long-term consequences of West Nile to a new biosecurity-level-3 laboratory at nearby Texas Children’s Hospital, where she will explore a link between the virus and kidney disease.

Researchers agree that the claim needs to be investigated. “If Murray’s findings are true, we have to think about what to do with all of these people with mild infections,” says William Reisen, an entomologist at the Center for Vectorborne Diseases at the University of California, Davis. But Murray is also facing scepticism, which she hopes to address in the latest phase of her research.

Murray’s quest began at a meeting of West Nile survivors in Texas in 2009, where a man in his early fifties who had recovered from a 2003 infection announced that he had kidney disease. He was dead within a year. To Murray, his illness brought to mind studies in which researchers had detected and cultured the virus in kidney tissue from laboratory animals long after they were infected with West Nile.

Murray collected urine samples from 25 survivors of West Nile and found that five had viral RNA in their urine well after they had been infected[1], suggesting that the virus might have established itself in their kidneys. To examine whether the virus might harm kidneys over time, Murray’s team then looked for indicators of long-term kidney disease, such as excess protein in the urine, in samples from 139 people, most of whom were infected with the 2003 strain of the virus. She reported[2] in July that 40% of that group showed signs of long-term kidney disease.

However, Lyle Peterson, director of the division of vector-borne infectious diseases at the CDC, sees no cause for alarm given the current evidence. He points out that Murray’s latest study did not include a control group, and that a CDC study of some West Nile survivors in Colorado found no evidence of viral RNA in the subjects’ urine[3].

Michael Busch, director of the Blood Systems Research Institute in San Francisco, California, told *Nature* that his team had also failed to find viral RNA, even though it had tested some of the same urine samples that Murray used. “Until multiple labs find the same results from the same blinded samples, everything must be taken with a grain of salt,” says Busch.

Murray maintains that there is an art to detecting RNA in urine. The single-stranded fragments are easily broken apart by enzymes in the fluid and by freezing and thawing when samples are shipped and stored. By contrast, samples that had travelled for just one hour from Murray’s lab to the University of Texas Medical Branch (UTMB) in Galveston for independent testing corroborated her

▶

MORE ONLINE

TOP STORY

MEUL/ARCO/NATUREPL.COM

Job swapping makes its mark on honeybee DNA go.nature.com/4isngf

MORE NEWS

- The astronomical unit gets fixed go.nature.com/sxqav1
- Human-rights court to rule on Costa Rica fertility-treatment ban go.nature.com/rufxfi
- Lunar gravity-mapping mission hints at thin crust go.nature.com/gn6spo

PODCAST

The brain never rests, our gut instinct to cooperate and electricity from heat go.nature.com/gww192

▶ findings[1]. Still, she says, she was shaken by the criticism. "I was starting to feel crazy. I wondered if I was committing career suicide."

Murray has regained her confidence. In one electron micrograph image, not yet published, she points to a group of fuzzy spheres that look like West Nile virus in a cell found in the urine of a woman who suffered an infection in 2003. Other particles in urine can look like small, fuzzy balls, critics note. But Murray hopes that within a year she will have evidence that will silence the sceptics. With her new lab and a four-year grant from the National Institutes of Health, she plans to recruit 440 people — half of whom had West Nile — to a study to look for kidney disease. By next month she also plans to try to isolate and grow the virus from urine.

Murray says it's not just a matter of defending her hypothesis, and others agree. "If we know this is a problem, we would know to monitor West Nile patients, and as soon as you saw some kidney enzymes acting abnormally, you could start to think about ways to prevent the progression of kidney disease," says Frederick Murphy, a virologist at the UTMB.

No drugs have been shown to be effective against West Nile, for either the immediate infection or the long-term effects. "We need to start thinking about how to treat it," Murray says. If her suspicions prove correct, that need will soon seem far more urgent than it did before. ■

1. Murray, K. *et al. J. Infect. Dis.* **201,** 2–4 (2010).
2. Nolan, M. S. *et al. PLoS ONE* **7,** e40374 (2012).
3. Gibney, K. B. *et al. J. Infect. Dis.* **203,** 344–347 (2011).

NANOTECHNOLOGY

Nano-safety studies urged in China

Exposure surveys and stronger regulations are required for the industry to thrive, researchers say.

BY JANE QIU IN BEIJING

Here is a recipe for anxiety: take China's poorly enforced chemical-safety regulations, add its tainted record on product safety and stir in the uncertain risks of a booming nanotechnology industry.

As an antidote to this uneasy mixture, the country should carry out more-extensive safety studies and improve regulatory oversight of synthetic nanomaterials, leading Chinese researchers said at the 6th International Conference on Nanotoxicology in Beijing this month. "This is the only way to maintain the competitiveness of China's nanotechnology sector," says Zhao Yuliang, deputy director of the Chinese Academy of Sciences' National Center for Nanoscience and Technology (NCNST) in Beijing. "We certainly don't want safety issues to become a trade barrier for nano-based products."

China's investment in nanotechnology has grown rapidly during the past decade, and its tally of patent applications in the field has surpassed those of Europe and the United States (see 'Patent boom'). But only 3% of the investment is used for safety studies, says Zhao, compared with about 6% of federal nanotechnology funding in the United States. "The situation must be changed soon," he says.

Nanoparticles — which measure from 1 to 100 nanometres in diameter — are chemically different from their corresponding bulk materials, and their potential toxicity can vary according to dozens of characteristics, such as size, surface area and coating.

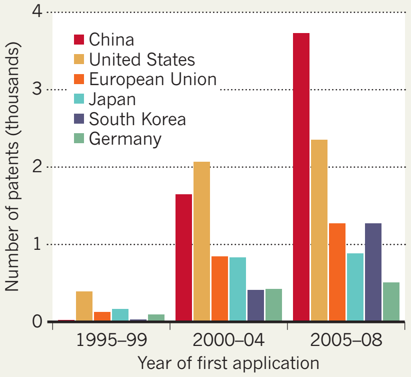
➲ NATURE.COM
read more from China at:
nature.com/nchina

In 2009, researchers claimed that nanoparticles were responsible for lung damage in seven workers at a printing factory in Beijing, two of whom subsequently died (see *Nature* **460,** 937; 2009). Volatile organic compounds may actually have been to blame, says Andre Nel, a nanotoxicologist at the University of California, Los Angeles, but such incidents could easily damage the fledgling industry's reputation. For now, however, the Chinese public remains unconcerned: in a survey led by Wang Guoyu, an ethicist at the Dalian University of Technology, nearly 80% of some 6,000 Chinese respondents said that they are not worried about the safety of nanoparticles.

Researchers at the meeting said that better safety testing was needed for products containing nanoparticles that can be absorbed by the body, such as food and cosmetics in which nanoparticles provide specific colours or textures. But occupational exposure among workers handling the materials may present the greatest risks: China's workplace safety rules are not always implemented, and they set no specific limits for handling nanoparticles.

PATENT BOOM

China has taken a global lead in filing nanotechnology patents over the past decade.



SOIURCE: PCAST/B. KISLIUK, USTPO. SEE GO.NATURE.COM/MQPWAU

"The main challenge is to tease out what characteristics make some nanoparticles hazardous," says Zhao. To address that question, Chinese researchers will next year join forces with colleagues in Europe, the United States and Brazil in a €13-million (US$17-million) project called Nanosolutions, to develop a nano-safety classification system based on material characteristics, toxicity studies and bioinformatics data. Initially focusing on 30 or so materials, such as carbon nanotubes, and nanoparticles of titanium dioxide and silver, the team will use high-throughput screening to identify the most toxic, and then investigate their biological effects in animal studies.

The data will be used to develop computer software to predict the potential hazards of other nanomaterials — a useful tool for industry and regulators around the world, and "essential to any progress in risk assessment", says Kai Savolainen, a researcher at the Finnish Institute of Occupational Health in Helsinki, who leads the project. "It will also help the industry to design safe nanomaterials from the outset."

Studying factory workers who are exposed to nanomaterials could yield further insights. With Chinese exposure levels likely to be much higher than those in Western factories, such surveys are ideally placed to quantify the risks involved, says Dhimiter Bello, an expert in occupational health at the University of Massachusetts in Lowell. Since last year, a team led by the NCNST's Chen Chunying has been monitoring chemical exposure levels, including those of nanomaterials, in three factories that have varying safety practices. The researchers hope that their data will help the government to draw up regulations covering nanoparticles in the workplace.

As China's exports are increasingly likely to carry nanotechnology inside them, better regulation is "important not only for safeguarding people, but for the public acceptance of nanotechnology worldwide", says Nel. ■

INFRASTRUCTURE

Laser centre lights up eastern Europe

European Union investment in high-energy-physics facility raises spirits in Romanian science community.

BY ALISON ABBOTT

Known for its cumbersome bureaucracy, the European Commission rarely makes people smile. But this week it has managed to gratify both a scientifically struggling former-communist country, and the world's nuclear-physics community.

On 18 September, the commission agreed to spend €180 million (US$237 million) on the first phase of construction for a futuristic nuclear-physics facility near Bucharest.

The Extreme Light Infrastructure Nuclear Physics Facility (ELI-NP) will generate laser pulses with up to 10 petawatts (10^{16} watts) of power, ten times the strength of current cutting-edge lasers — and intense enough to reveal the internal structures of atomic nuclei. "It will allow us to do a new sort of nuclear physics that hasn't been possible so far," says project leader Nicolae-Victor Zamfir, director-general of the Horia Hulubei National Institute of Physics and Nuclear Engineering in Măgurele, Romania, where the facility will be located. "The energy of the laser-light pulses will be almost at the level of the strong force that binds nuclei, so it will be able to perturb them."

Forty research institutions in 13 European Union (EU) member states have been involved in planning the facility, which is scheduled to start operating in 2017. The international academic community will be able to use the ELI-NP for free, but private companies will pay for access. Bids for access to the instruments will be assessed by international scientific committees.

The full construction cost for the facility will be €356.2 million, paid from Romania's allocation of structural funds — the EU subsidies designed to help poor regions to improve their infrastructure and economies. Structural funds have traditionally been used for civic projects such as road building, but the commission now encourages their use for projects that boost science.

The ELI-NP is one of three planned facilities in the Extreme Light Infrastructure, a broad effort to explore the frontiers of laser science that was identified as a top priority in 2006 by the European Strategy Forum on Research Infrastructure. All three will be built in eastern Europe, reflecting the commission's desire to balance the distribution of research infrastructures around the continent. The European Commission last year approved about €236 million for the ELI's first pillar in Prague, which will generate bursts of laser light in the 10-picosecond (10^{-13}-second) range to accelerate beams of particles to high energies so that their interaction can be studied. The third facility, which is planned for Szeged, Hungary, will produce even shorter radiation pulses, in the attosecond (10^{-18}-second) range, enabling physicists to image the dynamics of electrons in atoms, molecules, plasmas and solids.

The approval of the ELI-NP is a welcome confidence boost for scientists in Romania, which has one of the EU's lowest national investments in research — just 0.5% of gross domestic product, compared with an EU average of 2%. On top of this, Romanian scientists have this year been shamed by a series of high-profile plagiarism scandals (see *Nature* **488,** 264–265; 2012) and dismayed by the current government's restructuring of its research advisory councils to exclude members from abroad.

> **"Nothing can stop the facility from being built now."**

The ELI-NP will be insulated from politics, says Dragos Ciuparu, a chemical engineer at the Petroleum–Gas University in Ploieşti, Romania, who was secretary of state for research when the former government decided to commit the structural funds to the project two years ago. "So nothing can stop the facility from being built now," he says.

Ciuparu adds that the ELI-NP will help to keep researchers and engineers in Romania, which has suffered a major brain drain. "But future governments will have to invest seriously in the national physics community here to best reap the facility's advantages," he says. ■

CORRECTION

In the News Feature 'Dive master' (*Nature* **489,** 194–196; 2012), the text for the 'work basket' in the graphic should have read: "Carrying capacity has been doubled".

BURN OUT

BY MICHELLE NIJHUIS

Forests in the American west are under attack from giant fires, climate change and insect outbreaks. Some ecosystems will never be the same.

A little after noon on Sunday 26 June 2011, strong winds toppled an aspen tree onto a power line in the Jemez Mountains of northern New Mexico. The year had been extraordinarily dry, and the temperatures that week had soared well above normal. When a spark from the power line ignited a fire, wind gusts spread the flames into nearby dense stands of fir and pine.

Within an hour, ecologist Craig Allen, 55 kilometres away at his home in Santa Fe, learned about the fire in an e-mail from a US Forest Service fire manager. “I hope you guys catch this,” Allen wrote back. “We don’t need another big fire in the Jemez.”

But Allen could already see a plume of grey smoke rising to the west. By early the next morning, the Las Conchas fire had burned 17,500 hectares — about a hectare every three seconds. Within five days, it had grown to 42,000 hectares and become the largest fire in New Mexico’s history. By the time the fire was contained weeks later, it had burned more than 60,000 hectares of forest and scrubland, in many places roasting the vegetation so thoroughly that only charred stumps and bare dirt remained.

For Allen, who works for the US Geological Survey (USGS) in Los Alamos, New Mexico, and has been studying the forests of the Jemez Mountains for more than 30 years, a large fire was not unexpected, but its speed and intensity caught him off guard. Later that year, when Allen and other forest scientists toured the burned area, they were all stunned into silence. The size of the areas scorched bare by the fire dwarfed the patterns of past burns, judging from studies of fire scars in tree rings reaching back to 1600. This time, says Allen, the forest may not recover.

Across the American west, the area burned each year has increased significantly over the past several decades (see ‘Bigger Blazes’), a trend that scientists attribute both to warming and drying and to a century of wildfire suppression and other human activities. Allen suggests that the intertwined forces of fire and climate change will take ecosystems into new territory, not only in the American west but also elsewhere around the world. In the Jemez, for example, it could transform much of the ponderosa pine (*Pinus ponderosa)* forest into shrub land. “We’re losing forests as we’ve known them for a very long time,” says Allen. “We’re on a different trajectory, and we’re not yet sure where we’re going.”

DEAD ZONE

Thirteen months after the Las Conchas fire, on a clear, late-summer day, Allen hikes down a valley on the southern edge of the burn, just outside Bandelier National Monument, a tangle of canyons that shelter pre-Columbian cliff dwellings. Apart from a few shrubs and grasses, many of the hillsides — once covered

New Mexico's record 2011 fire left blackened stumps in place of evergreen pine and fir.

J. C. HONG/AP PHOTO; D. HANSON/ALBUQUERQUE JOURNAL/ZUMAPRESS.COM/ALAMY

with pine and fir — are empty and brown. Blackened stumps of alligator junipers (*Juniperus deppeana*), squat, gnarly trees that can live for many hundreds of years, stick up like twisted hands. Shade is just a wish and the landscape is still, with only a few raptors circling overhead. Allen is pleased to see some ants on the ground.

It looks as arid as Death Valley, but it is not. These mountains typically get more than 40 centimetres of rain a year, a healthy amount in the generally dry US southwest and enough — in theory — to regrow a forest. In the past, that is what happened. Because the fires burned unevenly, they left stands of surviving trees that then supplied seeds to scorched areas, allowing the forest to regenerate.

Over the past century, however, the policy of quickly dousing fires has allowed brush and spindly young trees to build up in many western forests, so they tend to burn hotter and less patchily than before. And over the past decade, a severe drought across the southwest has weakened trees and made them vulnerable to widespread attack by beetles, leading to a die-off of more than one million hectares of piñon pines (*Pinus edulis*)[1]. Many of the dead trees are still standing, and can serve as ladder fuels that transform relatively cool surface fires into hot, fast-moving crown fires that leap from treetop to treetop.

"It's not the area of these fires that's of most concern," says Allen. "It's the scale of the contiguous patches of dead trees."

Fires such as the Las Conchas one leave behind few seed sources, strip soils of nutrients and increase the likelihood of landslides. In their wake, vegetation of any kind can struggle to take root. When trees and shrubs do regrow, the region's warming temperatures and more frequent dry spells are likely to favour heat- and drought-tolerant species. By looking at tree rings, Park Williams of the Los Alamos National Laboratory and his colleagues have been able to assess how droughts stress southwestern forests[2]. They forecast that if temperatures rise as projected by climate models, trees will face worse drought stress in the first half of the twenty-first century than they have experienced for 1,000 years, probably driving a transformation of the ecosystem.

> "WE'RE LOSING FORESTS AS WE'VE KNOWN THEM FOR A VERY LONG TIME"

In some places in the Jemez, the transformation seems to have started. In 1996, the Dome Fire burned almost 7,000 hectares in the mountains, leaving patches of dead trees that at the time seemed surprisingly large, say Allen and others. Swathes of shrubby vegetation, dominated by scrub oaks, sprouted in the burned patches, surrounding small islands of surviving ponderosa pine and other conifers. When the Las Conchas fire roared through some of the same areas last summer, the oaks burned hot and fast, killing almost all the conifers that had survived the Dome fire.

Because the shrubs are better adapted to warmer, drier conditions than the trees, Allen expects that they will regrow in even larger patches. Eventually, they could dominate the entire landscape and establish a pattern of intense and frequent fires that is currently more common in coastal California and other Mediterranean-style ecosystems. On his hike through the burned area, Allen turns to Jorge Castro Gutiérrez, an ecologist at the University of Granada, Spain, who is visiting Santa Fe for the summer. "We're turning into something that's going to look very familiar to you," he says.

THE NEW NORMAL

All around the American west, scientists are seeing signs that fire and climate change are combining to create a 'new normal'. Ten years after Colorado's largest recorded fire burned 56,000 hectares southwest of Denver, the forest still has not rebounded in a 20,000-hectare patch in the middle, which was devastated by an intense crown fire. Only a few thousand hectares, which the US Forest Service replanted, look anything like the ponderosa-pine stands that previously dominated the landscape.

"Otherwise, it's grassland and shrub land, and probably will be for centuries to come," says Peter Brown, a forest ecologist and director of the non-profit organization Rocky Mountain Tree-Ring Research in Fort Collins, Colorado. From tree-ring analyses, he knows that even small bare patches left by crown fires in the nineteenth century have not returned to forest, so he holds little hope for the intensely burned patch from 2002.

In the Alaskan interior, as summers have turned warmer and drier and permafrost has thawed, fires are hitting more frequently and the fire season lasts longer. Burns reach deeper into soils and alter their chemistry. By favouring the regeneration of deciduous species, which are better adapted to burned soils and frequent fires, these changes could break what researchers call the 'legacy lock' of the black spruce (*Picea mariana*) forests in the region[3].

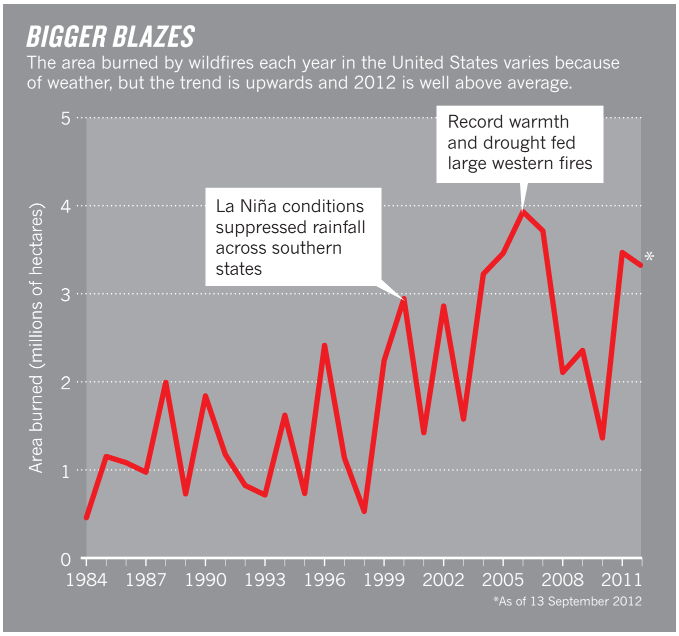
Similar dynamics are at work in the Great Basin and Sonoran deserts of the American west, where invasive grasses regenerate quickly after wildfires, creating a carpet of fast-burning fuel that makes future fires more likely. In the saguaro cactus (*Carnegiea gigantea*) forests of southern Arizona, wildfires were once an oddity. Now, flaming cacti are an increasingly common sight, and invasive species such as buffelgrass (*Cenchrus ciliaris*) are spreading.

In the high mountains of Glacier National Park, Montana, fire and warming may favour a new kind of forest — or none at all. Last year, Rachel Loehman of the US Forest Service's fire sciences lab in Missoula, Montana, and her colleagues used an ecosystem process model that simulated interactions between fire, climate and vegetation to assess the park's future. Their results[4] predict that climate change and frequent fires will trigger the spread of western white pine (*Pinus monticola*), a species common there until logging and wildfire suppression favoured the dominance of western red cedar (*Thuja plicata*), western hemlock (*Tsuga heterophylla*) and other shade-tolerant species. But as temperatures in the Rockies rise, western white pine is becoming vulnerable to white-pine blister rust (*Cronartium ribicola*) and beetle attack. If the western white pine succumbs to disease, the park could be left with little forest cover, says Loehman.

To keep the park's forests from disappearing, Loehman and her colleagues recommend that authorities there continue to conduct small 'prescribed' burns that limit future fires. She also supports ongoing efforts to develop and plant trees that have genetic resistance to blister rust.

Allen stressed these threats in August, when he testified before the US Senate Committee on Energy and Natural Resources about the issues being faced by western forests. Given the recent die-offs and the forecasts for the future, many forests could be heading towards a tipping point, he said. "If the climate projections of rapid warming for the Southwest are correct, then by the middle of the twenty-first century our Southwestern forests as we know them today will experience significant vegetation mortality and can be expected to reorganize with new dominant species."

The forest die-offs in the American west resemble shifts happening in other parts of the world. In 2010, Allen and 19 colleagues from around the world found that published reports of forest die-offs associated with drought have increased significantly since 1985, and are occurring in ecosystems ranging from the tropical forests of Costa Rica to Australian acacia forests and pine forests in east-central China[5]. They also found that no type of forest or climate zone was immune. With collaborators in Australia, Europe and throughout North America, Allen is now working to identify the physiological limits of various tree species, which should help in predicting future die-offs and changes in fire-prone areas worldwide. "We don't really understand what it takes to kill a tree," says Allen.



BIGGER BLAZES

The area burned by wildfires each year in the United States varies because of weather, but the trend is upwards and 2012 is well above average.

La Niña conditions suppressed rainfall across southern states

Record warmth and drought fed large western fires

*As of 13 September 2012

SOURCE: NATIONAL INTERAGENCY COORDINATION CENTER

FIXING FORESTS

Forest managers may try to slow down or stop the conversion of some forests, to preserve biodiversity, carbon storage or an iconic species, says Nathan Stephenson, a USGS ecologist at the Sequoia and Kings Canyon Field Station in Three Rivers, California, who studies one such icon — the sequoia. But that tends to require expensive and ongoing intervention, such as irrigating seedlings in habitats that are growing too hot and dry.

In other cases, managers will decide to let the forests shift, says Stephenson. "But the worst thing is for it to happen through disturbances that completely wipe out the vegetation, increase erosion and sap nutrients out of the soil," he says. "If we can find ways to ease the transition from one state to another, we'll be in much better shape."

One way to do that is to keep fires from spreading so quickly and burning so intensely. For well over a decade, the US Forest Service has been 'treating' some of its forest stands with selective logging and prescribed burning — not to prevent all fires, but to reduce the risk of large and severe burns. Managers at the Santa Fe National Forest in New Mexico, which includes the Jemez, now plan to treat about 45,000 hectares of forest. "One of the things we've learned is that these treatment areas have to be very large in order to work," says prescribed-fire specialist William Armstrong. That means not hundreds of hectares, but thousands or tens of thousands of hectares, he says.

Despite repeated urgings from Allen and other scientists, however, most treatment projects aren't on this scale. Funding and person power are often scarce, and the only way to thin large areas quickly and economically is to use prescribed burning — a tactic that generally meets with public resistance. "We're well aware of the science," says Armstrong, "and we're acutely aware that there's a whole lot we're not going to be able to do."

If large, high-severity fires do continue, some managers may choose to speed up landscape transitions, rather than slow them down. Nancy Grulke, director of the US Forest Service Western Wildland Environmental Threat Assessment Center in Prineville, Oregon, says that when fires burn through low-elevation conifer forests in the mountains around Los Angeles, California, and create landslide hazards, she advises managers not to replant conifers but to choose oaks and other species that are better adapted to a warm, dry future.

Given the uncertainties in how climate change, insect outbreaks and other stresses will affect forests in coming decades, Allen thinks that it is necessary to hedge bets after a fire by planting a range of species. He suggests building a "bridge to the future", by mixing some of the original tree types with species from lower elevations or warmer slopes, which could do well as conditions change.

That approach would help to make the ecosystems more resilient. But it will not restore the past, says Allen, who is saddened by the dramatic changes in the Jemez Mountains and beyond. At the end of a long, dry and fiercely hot hike through the Las Conchas burn, he surveys the bare hillsides and recalls what they were like just over a year ago — forested, cool and full of life. "For so many of us who have worked here for so long," he says, "this feels like a failure." ■

Michelle Nijhuis *is a writer in Colorado.*

1. Breshears, D. D. *et al. Proc. Natl Acad. Sci. USA* **102,** 15144–15148 (2005).
2. Williams, P. A. *et al. Nature Clim. Change* (in the press).
3. Wolken, J. M. *et al. Ecosphere* **2,** art124 (2011).
4. Loehman, R. A., Clark, J. A. & Keane, R. E. *Forests* **2,** 832–860 (2011).
5. Allen, C. D. *et al. Forest Ecol. Management* **259,** 660–684 (2010).

ILLUSTRATIONS BY RYAN SNOOK

IDLE MINDS

Neuroscientists are trying to work out why the brain does so much when it seems to be doing nothing at all.

BY KERRI SMITH

For volunteers, a brain-scanning experiment can be pretty demanding. Researchers generally ask participants to do something — solve mathematics problems, search a scene for faces or think about their favoured political leaders — while their brains are being imaged.

But over the past few years, some researchers have been adding a bit of down time to their study protocols. While subjects are still lying in the functional magnetic resonance imaging (fMRI) scanners, the researchers ask them to try to empty their minds. The aim is to find out what happens when the brain simply idles. And the answer is: quite a lot.

Some circuits must remain active; they control automatic functions such as breathing and heart rate. But much of the rest of the brain continues to chug away as the mind naturally wanders through grocery lists, rehashes conversations and just generally daydreams. This activity has been dubbed the resting state. And neuroscientists have seen evidence that the networks it engages look a lot like those that are active during tasks.

Resting-state activity is important, if the amount of energy devoted to it is any indication. Blood flow to the brain during rest is typically just 5–10% lower than during task-based experiments[1]. And studying the brain at rest should help to show how the active brain works.

➲ NATURE.COM
Listen to the *Nature Podcast* for more on the resting brain:
go.nature.com/gww192

Research on resting-state networks is helping to map the brain's intrinsic connections by showing, for example, which areas of the brain prefer to talk to which other areas, and how those patterns might differ in disease.

But what is all this activity for? Ask neuroscientists — even those who study the resting state — and many will sigh or shrug. "We're really at the very beginning. It's mostly hypotheses," says Amir Shmuel, a brain-imaging specialist at McGill University in Montreal, Canada. Resting activity might be keeping the brain's connections running when they are not in use. Or it could be helping to prime the brain to respond to future stimuli, or to maintain relationships between areas that often work together to perform tasks. It may even consolidate memories or information absorbed during normal activity.

"There's so much enthusiasm about the approach now, and so little basic understanding," says Michael Greicius, a neuroscientist at Stanford University in California, who started studying resting-state networks a decade ago.

ALWAYS ACTIVE

A set of experiments in the mid-1990s first suggested that the brain never really takes a break. Bharat Biswal, then a PhD student at the Medical College of Wisconsin in Milwaukee, was trying to find ways of identifying and removing background signals from fMRI scans, in the hope that it would improve interpretations of the signals from tasks. "The assumption was, it was all noise," says Biswal, who is now a biomedical engineer at the New Jersey Institute of Technology in Newark. But when he looked at scans taken when people were resting in the scanner, he saw regular, low-frequency fluctuations in the brain[2]. Biswal's experiments suggested that neuronal activity was causing these fluctuations.

In the early days of resting-state research, some people were sure that they had found something profound. "When I first started looking at these networks, I was convinced we were tapping into the stream of consciousness, and this was real-time ongoing conscious processing," says Greicius. But, he says, "I was relatively quickly disabused of that notion". The networks of activity also appeared in altered states of consciousness such as when sleeping or under anaesthesia[3,4], so they weren't necessarily linked to conscious processing.

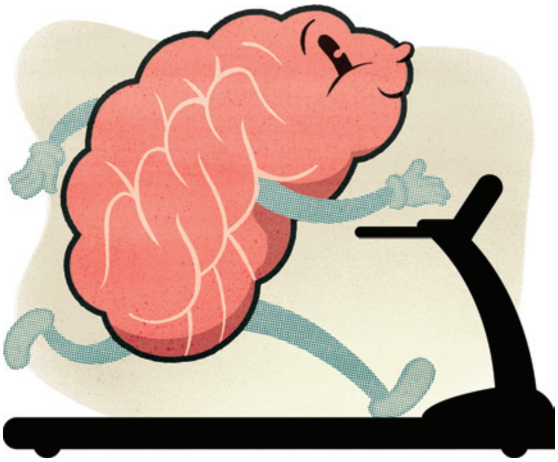
But they weren't meaningless either. Several years after Biswal's discovery, studies of the resting state in its own right began to emerge. A team led by Marcus Raichle, a neuroscientist at Washington University in St. Louis, Missouri, characterized[5] activity in one such network as the brain's default mode — what they considered its baseline setting. During tasks, default-mode activity actually dropped, coming back online when the brain was no longer focusing so intensely[5].

The default-mode network has been joined by dozens of other flavours of resting-state network — some of which resemble the circuitry that contributes to attention, vision, hearing or movement. They seem very similar across study participants but are also dynamic, changing over time. "The fact that it's always present but modifiable tells you that it's got its importance," says Michael Milham, director of the Center for the Developing Brain at the Child Mind Institute in New York.

Still, some researchers have questioned whether these resting patterns represent anything real. After all, fMRI does not measure brain electrical activity directly: it monitors blood flow. The low-level idling activity could be an artefact. "People suspected it was lousy scanners or respiratory noise," says Andreas Kleinschmidt, director of research at the French National Institute of Health and Medical Research's Cognitive Neuroimaging Unit in Gif-sur-Yvette. But using fMRI and electroencephalography (EEG) recordings, Kleinschmidt and his team confirmed[6] that various resting-state networks are correlated with real neural activity.

Shmuel and David Leopold, a neurophysiologist at the US National Institute of Mental Health in Bethesda, Maryland, did much the same[7], imaging resting states in monkeys while recording the animals' electrical brain activity using probes implanted deep in the visual cortex. They found correlations between resting-state networks and electrical activity in a band of frequencies around 40 hertz. Such 'γ activity' is associated with communication between distant brain areas, and seeing it convinced Shmuel that resting-state networks represent actual brain activity. "I strongly believe that there is a neurophysiological mechanism that underlies the entire thing that we call resting-state networks," he says.

RESTING-STATE ACTIVITY IS IMPORTANT, IF THE AMOUNT OF ENERGY DEVOTED TO IT IS ANY INDICATION.



DISORDERED THINKING

It is a mechanism that may go awry in brain disorders. People with early signs of Alzheimer's disease, for example, have unusual resting-state signatures that can be detected even at very mild levels of dementia and which vary as the disease progresses[8]. In children with autism spectrum disorder, resting-state networks can be 'hyperconnected', displaying more links than for kids without the condition[9]. The reasons for these differences are not clear, and they may not matter to clinicians, who are interested in finding disease markers. "From a clinical perspective, you're not always going to understand why a biomarker is serving as that biomarker," says Milham. But some neuroscientists are deeply curious as to what these fluctuations do. "It keeps me up at night," says Timothy Ellmore at the University of Texas Health Science Center in Houston, who is studying resting brain activity in people with Parkinson's disease.

Some researchers now think that resting-state networks may prime the brain to respond to stimuli. "The system is not sitting there doing nothing and waiting," says Kleinschmidt. Cycling activity in these networks may be helping the brain to use past experiences to inform its decisions. "It's incredibly computationally demanding to calculate everything on the fly," says Maurizio Corbetta at Washington University School of Medicine in St. Louis. He has been studying resting state using magnetoencephalography, a technique that measures magnetic fields associated with the electrical activity of neurons. "If I have ongoing patterns that are guessing what's going to happen next in my life, then I don't have to compute everything." He likens the activity to the idling of a vehicle. "If your car is ready to go, you can leave faster than if you have to turn on the engine."

"IF YOUR CAR IS READY TO GO, YOU CAN LEAVE FASTER THAN IF YOU HAVE TO TURN ON THE ENGINE."

But idling networks might not just save time. They may also influence perceptions — albeit unconsciously. To study how spontaneous resting activity affects perception, Kleinschmidt and his colleagues scanned[10] the brains of people who were looking at a picture that can be perceived as a face or as a vase. Study participants who reported seeing a face had more spontaneous activity in the fusiform face area — a brain region that processes faces — before they were shown the picture. Kleinschmidt suspects that the brain is running several models of the world in the background, ready for one of them to turn into reality. "Ideally, you're always prepared for what happens next," he says.

Corbetta has discovered evidence in people with brain damage that resting activity can change behaviour. In unpublished work, he has found hints that lesions in frontal brain regions — caused by stroke, for example — can give rise to changes in spontaneous brain activity in distant areas. What is more, the changes to the resting activity are linked to the behavioural deficit. "This is clear evidence that resting-state impairments are affecting the way the network is recruited during a task," he says.

ZEN AND THE ART OF NETWORK MAINTENANCE

Raichle favours the idea that activity in the resting state helps the brain to stay organized. The connections between neurons are continually shifting as people age and learn, but humans maintain a sense of self throughout the upheaval. Spontaneous activity might play a part in maintaining that continuity. "Connections between neurons turn over in minutes, hours, days and weeks," says Raichle. "The structure of the brain will be different tomorrow but we will still remember who we are."

Or perhaps the activity is part of the reshaping process, tweaking connections while we idle. Several teams have reported changes in resting connectivity after language and memory tasks and motor learning. Chris Miall, a behavioural brain scientist at the University of Birmingham, UK, and his colleagues have shown that spontaneous activity at rest can be perturbed by what has just happened[11]. The team scanned volunteers at rest, and then asked them to learn a task involving using a joystick to track a moving target. When the participants were scanned at rest again, the team could see the effects of motor learning in the resting networks. That study, and subsequent work along the same lines, suggests that "the brain is not only thinking about supper coming up, but it's also processing the recent past and converting some of that into long-term memories", says Miall. The network changes are specific to the tasks performed.

Work on memory consolidation in animals backs that conclusion. It used to be assumed that memories from the day were strengthened during a night's sleep. Working with rats, however, Loren Frank and Mattias Karlsson, neuroscientists at the University of California, San Francisco, have found[12] that the brain replays and consolidates new memories at any chance it gets — even when awake. "These events happen when it doesn't look like the animal is doing very much," says Frank.

He speculates that resting activity could be doing the same thing in human brains — reactivating patterns that correspond to past experiences. At the same time, activity in the networks could have a normalizing, housekeeping function too. "How do you keep the brain flexible?" Frank asks. "If you have random patterns of activity washing through your network, those can help reduce the strength of the pathways associated with what you've just learned." That would stop the brain from reinforcing the same pathways too often. "Perhaps down-time periods are also important for that," he says.

Shmuel says that it is still not possible to rule out the idea that this activity is just a by-product of the brain being alive. Current may flow through these circuits "simply because there is current — the brain is not dead — and there are anatomical connections that give this current a non-random structure". But, he admits, "I hope this is not the case. Then it's extremely uninteresting."

Narrowing down the range of interesting possibilities may take time, given that the very nature of resting-state science makes it difficult to test hypotheses. When a researcher slides someone into a scanner and instructs them to think about nothing in particular, there is no task to do and no hypothesis to address. So researchers have to generate reams of data and line up hypotheses as they go along. "Resting state opens up discovery science," says Milham enthusiastically, before admitting that, because he trained as a hypothesis-driven cognitive neuroscientist, "it's like heresy that I've got into this".

Whatever resting activity is doing, its existence certainly proves one thing. Miall puts it bluntly: "The brain only rests when you're dead." ■

Kerri Smith *is podcast editor for* Nature *in London.*

1. Raichle, M. E. & Mintun, M. A. *Annu. Rev. Neurosci.* **29,** 449–476 (2006).
2. Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. *Magn. Reson. Med.* **34,** 537–541 (1995).
3. Greicius, M. D. *et al. Hum. Brain Mapp.* **29,** 839–847 (2008).
4. Boly, M. *et al. Ann. NY Acad. Sci.* **1129,** 119–129 (2008).
5. Raichle, M. E. *et al. Proc. Natl Acad. Sci. USA* **98,** 676–682 (2001).
6. Laufs, H. *et al. Proc. Natl Acad. Sci. USA* **100,** 11053–11058 (2003).
7. Shmuel, A. & Leopold D. A. *Hum. Brain Mapp.* **29,** 751–761 (2008).
8. Brier, M. R. *et al. J. Neurosci.* **32,** 8890–8899 (2012).
9. Di Martino, A. *et al. Biol. Psychiatry* **69,** 847–856 (2011).
10. Hesselmann, G., Kell, C. A., Eger, E. & Kleinschmidt, A. *Proc. Natl Acad. Sci. USA* **105,** 10984–10989 (2008).
11. Albert, N. B., Robertson, E. M. & Miall, R. C. *Curr. Biol.* **19,** 1023–1027 (2009).
12. Karlsson, M. P. & Frank, L. M. *Nature Neurosci.* **12,** 913–918 (2009).

COMMENT

ASTRONOMY Call for the US to rejoin the Square Kilometre Array telescope **p.363**

BIOWARFARE Behind the scenes of the USSR's bioweapons programme **p.364**

ART CERN's artist-in-residence on why he whips trees **p.366**

OBITUARY Neil Armstrong, first human on the Moon, remembered **p.368**

JIM RICHARDSON

Rhoda Mang'yana grows maize near 'fertilizer trees' to improve her farm's crop yield and soil fertility.

Plant perennials to save Africa's soils

Integrating perennials with food crops could restore soil health and increase staple yields, say **Jerry D. Glover**, **John P. Reganold** and **Cindy M. Cox**.

Rhoda Mang'yana's half-hectare farm in Malawi produces more maize (corn) than her extended family of seven can eat. Some of the extra crop she sells. Some she feeds to pigs and goats, which she also sells. With the money she makes, she can pay her grandchildren's school fees and buy essentials, such as soap and salt, that she has provided for her family since her husband died 15 years ago. As well as maize, Mang'yana's farm supplies firewood and other types of animal feed. It is also resilient, providing enough maize during good years to pull the family through leaner ones. Key to Mang'yana's improved land is perenniation — the integration of trees and perennials (plants that live for two or more years) into fields of food crops.

When Mang'yana acquired the farm in the 1990s, soil degradation limited its annual maize yield to less than 1 tonne per hectare — a common yield in Africa, but one-tenth of those seen in the Corn Belt of the US Midwest. To improve the soil, she began growing perennial pigeon peas (*Cajanus cajan*) and groundnuts (*Arachis hypogaea*), which require less fertilizer and add nitrogen to the soil[1]. She also began using 'evergreen agriculture', planting various nitrogen-fixing trees[2] that each meet different needs. Fast-growing, short-lived plants such as *Gliricidia sepium* provided firewood and animal feed; slower-growing, longer-lived trees such as *Faidherbia albida* improved long-term soil fertility.

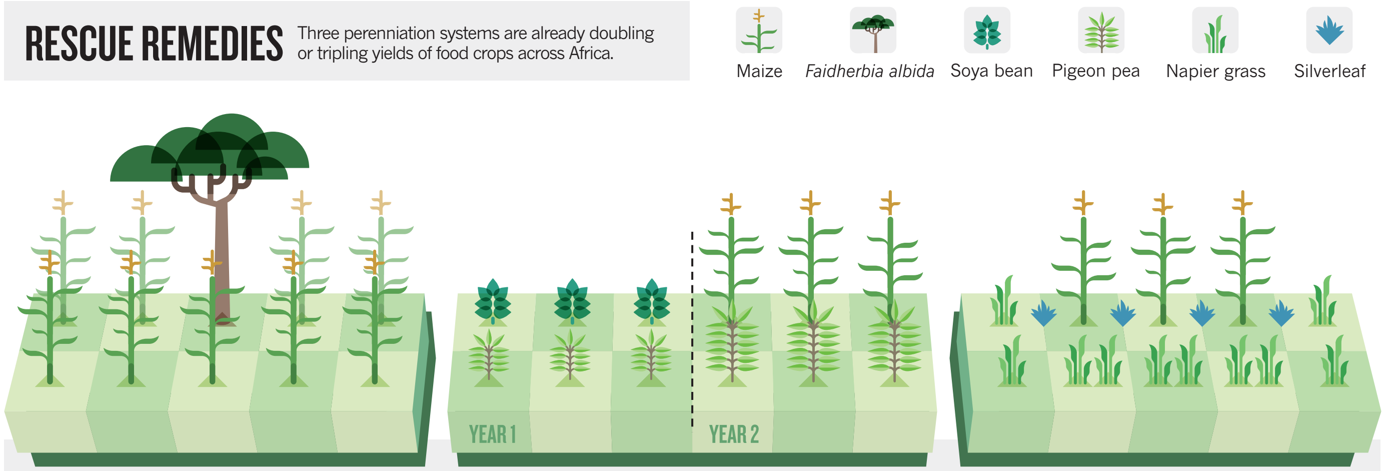
After a few years, Mang'yana resumed growing maize. Better yields allowed her to invest in pigs and goats, and she began using the animals' manure along with mineral fertilizer on the fields. Now she produces four tonnes of maize per hectare in a good year. Most of Africa's soils are naturally poor in nutrients compared with those of the younger landscapes found in North America, for example. Only about 16% of Africa's lands have the high-quality soils best suited to supporting livestock and crops; roughly 29% are marginal; and the rest are unsuitable for farming[3]. Farmers often worsen already poor lands by removing more nutrients than they return to the soil[4].

Population growth and erratic weather driven by climate change are exacerbating the problem[5]. Although cereal production grew by 2% a year in most African countries

▶

RESCUE REMEDIES

Three perenniation systems are already doubling or tripling yields of food crops across Africa.



EVERGREEN AGRICULTURE

- Improves yields and provides feed, firewood
- Maintains vegetative soil cover year-round
- Increases available water and nutrients in soil
- Enhances soil carbon stores

DOUBLED-UP LEGUME SYSTEM

- Increases plants' efficiency of fertilizer use
- Improves yield of protein-rich grains
- Decreases labour requirements
- Improves families' diets

PUSH–PULL

- Helps to suppress insect pests and weeds
- Reduces need for external inputs
- Provides animal feed
- Reduces soil erosion

▶ during 1961–2003, the population grew faster (2.6% annually), leading to an overall 43.5% decline in per capita cereal production over that period[6].

About one-quarter of the world's undernourished population — some 240 million people — live in sub-Saharan Africa. Of the various factors needing urgent attention to increase agricultural productivity, scientists from the region have identified soil quality as a top priority[5]. We believe that perenniation should be used much more widely to help farmers to meet the challenge of improving soils while increasing food production.

DEEP ROOTS

Many African farmers struggle to meet the nutrient needs of their crops. Because organic sources such as animal and plant manure are often in short supply in Africa, governments and development agencies tend to promote mineral fertilizers as the solution to low soil fertility. But investing in fertilizer can be risky — during a drought year, for instance, farmers might not produce enough to cover the costs. And without organic inputs, mineral fertilizers do little to improve soil conditions, and can even worsen them by hastening the loss of soil carbon[7].

Perennials can gain access to more of the soil's nutrients and water, for a longer time than annual crops. Their roots often extend more than two metres deep (compared with less than a metre for most annuals), and their growing seasons are longer. These attributes make them more resilient to harsh environmental conditions. Because they produce more biomass, both above and below ground, they are better at reducing soil erosion, transferring organic inputs to soil microorganisms and increasing the amount of carbon stored in the soil — a key component of soil health[8]. These organic inputs and microorganisms then improve soil fertility and structure as well as increase water infiltration and storage — all of which increase the amount of water available to and used by crops[7,8]. Moreover, by supplying the soil with carbon, perennials can improve the ability of food crops to use mineral fertilizers and, potentially, help farmers to adapt to climate change[1,2,8].

Here we highlight three perenniation approaches that show particular promise in sub-Saharan Africa: evergreen agriculture and doubled-up legume systems, such as those used by Mang'yana, and a technique for managing crop pests called 'push–pull' (see 'Rescue remedies').

> **"Without organic inputs, mineral fertilizers can worsen soil conditions."**

Evergreen agriculture is the best known and most widely adopted of the three. Hundreds of thousands of farmers across the Sudano-Sahelian zone and into East and Southern Africa grow 'fertilizer trees' along with maize, sorghum or millet on more than 5 million hectares of cropland[2]. The leguminous trees in these systems, such as *Faidherbia albida*, can triple maize yields while improving the soil. The trees meet other needs as well — they provide firewood and livestock fodder, for example[2].

In doubled-up legume systems, which have now been adopted by more than 8,000 households in Malawi[1], farmers grow perennial pigeon pea along with annual legumes such as soya beans (*Glycine max*) or groundnuts. After harvesting the legumes, farmers plant maize in or beside the rows of pigeon peas and then harvest both. Farmers can use different types of pigeon pea, depending on how much grain they need for food and leaves and stems for animal feed or manure. They can also change the timing and arrangement of planting to favour the maize or the legume. Nutrient- and protein-rich, pigeon peas can persist into the drier months, after maize stocks have been exhausted[9], and they can substantially improve families' diets.

Perennial plants can help to manage pests and diseases. More than 30,000 farmers in East Africa have adopted push–pull systems to manage stem-borer moths (*Chilo partellus*) and African witchweed (*Striga hermonthica*), both widespread in sub-Saharan Africa. In this method, silverleaf (*Desmodium uncinatum*), a perennial leguminous animal-feed crop, is interspersed among maize plants. The silverleaf produces chemicals that repel or 'push' pests away, and perennial Napier grass (*Pennisetum purpureum*) grown around the edges of the fields 'pulls' the pests in by providing attractive leaves for egg-laying. Push–pull systems can more than double maize yields by reducing pests[10] and increasing the amount of nitrogen in the soil.

Each of these three soil-building systems can be adapted to specific types of farming, such as conservation agriculture, organic or conventional management or production of both crops and livestock.

Organizations such as the US Agency for International Development (USAID)

JIM RICHARDSON

Mang'yana's pigs, which she sells to pay her grandchildren's school fees, eat maize bran, *Gliricidia* branches and weeds that grow in the fields.

and the World Bank have made sustained investments in strategies discovered and developed by farmers, and these efforts have proved crucial in battling hunger over the past 50 years. Irrigation and fertilization have become specialized scientific disciplines, sparking the creation of dedicated research institutes around the world.

In many regions of Africa, farmers have taken some perenniation approaches well beyond the proof-of-concept stage. Yet many questions remain — such as which species are best suited to which types of land, and how to maximize productivity in different areas. We believe that perenniation, along with technologies such as improved seed, fertilization and irrigation, should be a priority for the international agricultural research-and-development community. This means scaling up the use of approaches known to work, such as evergreen agriculture (in suitable areas), and backing research in cultivars and techniques that farmers have not yet tested widely.

SCALING UP

Some efforts to expand perenniation are already under way. Last month, a four-year project called Trees for Food Security was launched by the World Agroforestry Centre, an international research institute based in Nairobi, Kenya, that has led the development of evergreen agriculture. The centre aims to plant millions of trees on farmland throughout sub-Saharan Africa, in partnership with the governments of Ethiopia, Rwanda, Burundi and Uganda.

Similarly, the International Crops Research Institute for the Semi-Arid Tropics, based in Patancheru, India, has worked for more than two decades with pigeon peas, collecting and characterizing cultivars and educating farmers about their use. The institute's collaboration with Michigan State University and others has boosted the use of doubled-up legume systems considerably over the past 10 years, particularly in Malawi[1].

Many research institutes, including Washington State University in Pullman[8], have taken up the development of perennial grains more broadly. And USAID is investing US$9 million annually in Africa Research in Sustainable Intensification for the Next Generation, a programme that includes support for the study of perenniation strategies (www.africa-rising.net). Yet these are drops in the ocean compared to the scale of need.

"Many farms have taken perenniation well beyond the proof-of-concept stage."

Giving perenniation the kind of support now directed towards technologies such as mineral fertilizers and seed development could require hundreds of millions of dollars. According to Chris Reij, an expert in African agriculture at the World Resources Institute in Washington DC, $50 million would be needed even to "show how existing successes [in agroforestry] could be scaled up".

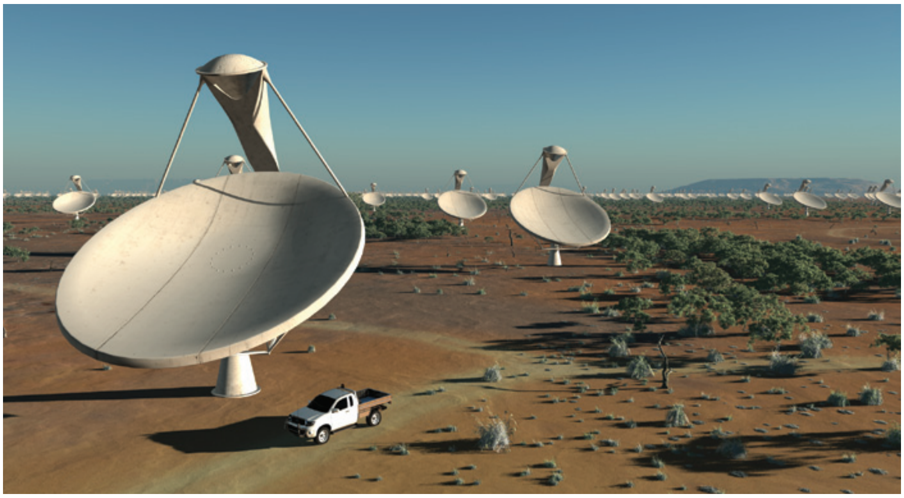
Yet such numbers pale in comparison to the losses of nitrogen, phosphorous and potassium from sub-Saharan farm fields each year, which are estimated to be equivalent to billions of dollars in fertilizer[4].

Sub-Saharan Africa's population is expected to reach 1.5–2 billion by 2050. Already the population is ballooning; in many areas, the risk of drought and flood is increasing; most soils are poor; and richer nations are buying up Africa's arable land for their own food or fuel security. African farmers have demonstrated the promise of perenniation. It is time to scale up its use and put it firmly on the research-and-development map. ■

Jerry D. Glover *is with the USAID Bureau for Food Security, Washington DC 20523, USA.* **John P. Reganold** *is in the Department of Crop and Soil Sciences, Washington State University, Pullman, Washington 99164, USA.* **Cindy M. Cox** *is at the International Food Policy Research Institute, Washington DC 20006, USA.*
e-mail: jglover@usaid.gov

1. Snapp, S. S., Blackie, M. J., Gilbert, R. A., Bezner-Kerr, R. & Kanyama-Phiri, G. Y. *Proc. Natl Acad. Sci. USA* **107,** 20840–20845 (2010).
2. Garrity, D. P. *et al. Food Sec.* **2,** 197–214 (2010).
3. Eswaran, H., Almaraz, R., van den Berg, E. & Reich, P. *Geoderma* **77,** 1–18 (1997).
4. Henao, J. & Baanante, C. *Agricultural Production and Soil Nutrient Mining in Africa: Implications for Resource Conservation and Policy Development* (International Center for Soil Fertility and Agricultural Development, 2006).
5. Committee on a Study of Technologies to Benefit Farmers in Africa and South Asia, National Research Council. *Emerging Technologies to Benefit Farmers in Sub-Saharan Africa and South Asia* (National Academies Press, 2008).
6. Betru, S. & Kawashima, H. *Afr. J. Agric. Res.* **5,** 2757–2769 (2010).
7. Bationo, A., Kihara, J., Vanlauwe, B., Waswa, B. & Kimetu, J. *Agr. Syst.* **94,** 13–25 (2007).
8. Glover, J. D. *et al. Science* **328,** 1638–1639 (2010).
9. Adu-Gyamfi, J. J. *et al. Plant Soil* **295,** 127–136 (2007).
10. Hassanali, A., Herren, H., Khan, Z. R., Pickett, J. A. & Woodcock, C. M. *Phil. Trans. R. Soc. B* **363,** 611–621 (2008).

Disclaimer: the views expressed in this article are not necessarily those of USAID.



SKA ORG./TDP/DRAO/SWINBURNE ASTRONOMY PRODUCTIONS

Artist's impression of part of the Square Kilometre Array, which will be sited in South Africa and Australia.

The United States must rejoin the SKA

Plans for the radio-telescope array must be firmed up to help Americans get back on board and ensure its success, say **Anthony J. Beasley** and **Ethan J. Schreier**.

The world's most powerful astronomical instrument is currently being built in South Africa and Australia. A growing consortium of countries, soon to be marshalled by incoming director-general Phil Diamond, is laying the foundation for some unique science. When it comes online next decade, the Square Kilometre Array (SKA) will observe diffuse hydrogen ionized by the first stars and galaxies, use pulsars to explore general relativity, and detect the imprints of dark energy on the distribution of matter in the Universe.

There is one country notable by its absence in this endeavour: the United States. And its absence threatens to hinder the SKA's pursuit of its scientific goals. After nearly 20 years of participation, US astronomers last year dropped out of the SKA collaboration as the result of disillusionment with the project's planning process and budget pressure from the National Science Foundation (NSF).

This cannot be allowed to continue: the United States must eventually rejoin the SKA. We call on the NSF to plan for a US role in the SKA, and we urge the SKA consortium and Phil Diamond to review the programme's goals and produce a realistic plan for achieving them.

In late 2011, the NSF ceased to fund any US participation in SKA development. This blow could be compounded if the NSF adopts its panel's recommendation to stop supporting — and so potentially close — the Green Bank Telescope and the Very Long Baseline Array (VLBA), both state-of-the-art and cost-effective telescopes with which SKA technologies could be evaluated.

Our experience in building large telescopes on the ground and in space leads us to believe that these decisions are short-sighted. They leave US astronomers and engineers unable to contribute to the SKA design or to participate in its science. The global astronomy community will press ahead without the United States. But without US scientific and technical input, and the ability to test SKA technologies at our facilities, the array's development will be slowed down by many years.

A lack of clarity on technical details and costs were the main criticisms of the SKA in the US astronomy community's 2010 decadal survey (go.nature.com/4qyqle), which considered the project scientifically exciting but only partly defined. We agree. Satisfying all the telescope's ambitious goals will require several different types of technology (such as receiving dishes, dipoles and tiles) and the consortium has yet to decide how to adapt and integrate them. Participation of US astronomers will be crucial in the firming up of those plans.

The scientific community recognizes that seed funding and development work towards the next generation of facilities is important, and that gaps in funding only add cost and delay. The NSF should continue to support the operation of existing radio-astronomy facilities in the United States, maintaining core capabilities that will also be necessary as test-beds for SKA technology in the coming decade. The US$10 million to $15 million per year needed to retain Green Bank and the VLBA is small relative to the billions already invested in US radio astronomy, which draws upon one-third of the NSF's annual $230 million astronomy budget.

THE WAY FORWARD

NSF funding for SKA development should be re-established. Even low initial levels (around $100,000 per year) would support planning activities, travel to meetings and some basic technology research. US facilities and university astronomy groups should together develop a strategy for participating in SKA planning and prototyping. By 2015, the United States should rejoin the SKA as a full partner.

In the next 1–2 years, Phil Diamond and the consortium should decide the technical requirements for the SKA (including frequency range, field of view, angular resolution and sensitivity), and should clearly define the technology developments necessary for a realizable instrument. This will be tricky because the ambitious goals of the SKA hinge on continual 'Moore's law' improvements in digital technologies.

Components available now will be obsolete by the time the telescope comes online, so a gradual updating process needs to be worked into the plan. Planners must project what technologies will become available in the next decade, and pick those that are feasible within a reasonable funding envelope.

Progress must be synchronized with projections of funding, so that partner contributions can be integrated steadily. To maintain project momentum, detailed design and development efforts should be paced and not completed long before construction money becomes available.

With US involvement and careful planning, the global radio-astronomy community can drive an evolutionary path towards the SKA, one that builds on current investments while enabling major discoveries as we advance. ■

Anthony J. Beasley *is director of the National Radio Astronomy Observatory.* **Ethan J. Schreier** *is president of Associated Universities Incorporated.*
e-mail: tbeasley@nrao.edu

CORRECTION

The article 'Beyond the Higgs' (*Nature* **488,** 581–582; 2012) located the RENO experiment in Seoul instead of Yonggwang.

Soviet maps omitted the 'Progress' plant in Stepnogorsk, once the world's largest bioweapons facility.

DAVID HONL/ZUMA PRESS

MILITARY SCIENCE

The USSR's deadly secret

Tim Trevan weighs up an authoritative take on the Soviet Union's vast, covert and costly bioweapons programme.

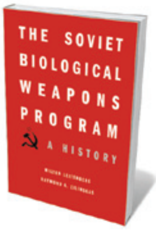
Two key events in the history of biological weapons occurred in 1972. The Biological and Toxin Weapons Convention was signed, with the United States, Britain and the Soviet Union as depositaries, or administrators. At the same time, in blatant violation of that convention, the Soviets re-energized their bioweapons programme, launched in the aftermath of the First World War. This massive, covert research push was the only biowarfare programme known to have modified pathogens through genetic engineering.

Conducted in great secrecy over two decades, the programme cost billions of rubles and involved up to 65,000 scientists and technicians. Some worked in the military; many others in a network of civilian laboratories under 'legends', or multilayered cover stories. The programme ended only after microbiologist Vladimir Pasechnik defected to Britain in October 1989. Milton Leitenberg and Raymond Zilinskas explore this murky world exhaustively in *The Soviet Biological Weapons Program*.

The book is peerless as a reference on the Soviet bioweapons programme, and highlights areas where not enough is known and worries remain. No page-turner, this is a densely factual, acronym-laden and footnoted catalogue of open-source and interview material gathered during more than ten years of meticulous research. The notes alone are a major contribution to the field.

Leitenberg and Zilinskas chronicle the decision-making process behind the programme, as well as its achievements and failures. They examine the performance of US and UK intelligence and diplomatic services in failing to uncover what was going on before Pasechnik's defection, and subsequently in negotiating the Trilateral Accords — the consultations between the Russian Federation, United States and Britain about his allegations.

The book also provides fascinating insight into how vested interests in both the Soviet military and its scientific establishment thwarted the decision to shut down the offensive parts of the programme by the Central Committee in 1989, Mikhail Gorbachev in 1991 and Boris Yeltsin the following year. And it throws some light on the current debates over H5N1 and other dual-use research — work with the potential to be used both for beneficial and malicious means.



The Soviet Biological Weapons Program: A History
MILTON LEITENBERG AND RAYMOND A ZILINSKAS
Harvard University Press: 2012. 960 pp. £40.95, $55

In the context of Soviet warfare, the programme had a handful of signal achievements. For example, it created strains of bacteria resistant to multiple antibiotics. It also genetically altered *Bacillus anthracis* (anthrax) to render existing vaccines ineffective. Perhaps most alarmingly, it genetically altered *Legionella pneumophila* (the agent responsible for Legionnaires' disease) to precipitate an immune-system attack on myelin, the main insulating material in the human nervous system. Such attack creates an artificial, rapid-onset disease similar to multiple sclerosis.

According to the authors, the bioweapons programme achieved little in terms of defences against pathogens. They also write that no systems existed to deliver biological warfare to the continental United States, and that work on intercontinental ballistic missile and cruise missile warheads never progressed far — an assessment that may raise eyebrows in the Western intelligence community.

What does this book's depiction of the Soviet genetic-engineering effort teach us about the dangers of new threats, using even more advanced techniques and vastly greater knowledge?

First, it debunks the theory of biological warfare as the 'poor man's nuke', at least as a weapon of mass destruction (as opposed to one of terror or assassination). The

Signs hint at the dangers that were once prevalent at the USSR's Progress bioweapons plant.

DAVID HONL/ZUMA PRESS

programme soaked up staggering amounts of money, expertise, resources and time.

Second, it proves that it is technically very difficult to genetically engineer pathogens to meet all 12 of the criteria for military usefulness, such as being suitable for aerosol delivery and able to survive in stable form in the air. Pleiotropy — the fact that a single gene may affect more than one feature in an organism — often means that efforts to enhance one 'desirable' property reduce others. However, advances in genomics may eventually overcome this obstacle.

Third, it shows that there was no national strategy for the programme. As Leitenberg and Zilinskas note, it did not benefit the Soviet Union's ability to wage war, but it did severely impair economic development in biotechnology by diverting scientific talent. Indeed, there was no stated doctrine of use laying out in what circumstances or how the Soviet armed forces would use bioweapons on the battlefield.

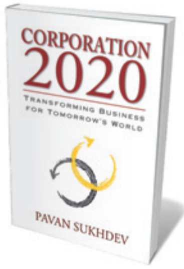
Finally, the book illustrates the impracticality of applying a single-use/dual-use approach to biology. Rather, we should talk of use and misuse. Even the most apparently 'single-use' aspect of the Soviet offensive research — the development of antibiotic-resistant pathogens — has potentially large-scale health applications. An attenuated, antibiotic-resistant live vaccine could be injected into patients with a disease and under antibiotic treatment. The antibiotics would attack the disease, but not the vaccine. We should seek not to ring-fence 'dual-use' technologies — impossible, in any case — but to discover how to prevent the misuse of biology.

Biopreparat, the ostensibly civilian part of the programme, came about because Soviet military bioweapons experts wanted parity with nuclear experts; at the same time, civilian scientists realized that their research would be funded only if it had weapons applications. As the programme's scientific champion Yury Ovchinnikov is reported to have said: "Nobody would give us money for medicine. But offer one weapon and you'll get full support."

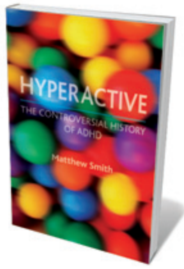
In other words, because Soviet biologists were underfunded and under-respected, they distorted what they offered to fit military funders' misinformed biases. Behavioural economists tell us how decisions that are logical at the individual level can result in outcomes that are, at the aggregate level, wildly illogical. Perhaps the Soviet programme, a clear example of this, tells us that behavioural economists should have a role in analysing how to prevent proliferation or create environments conducive to disarmament. ■

Tim Trevan *is the executive director of the International Council for the Life Sciences in McLean, Virginia.*
e-mail: trevan@iclscharter.org

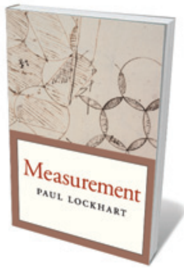
Books in brief



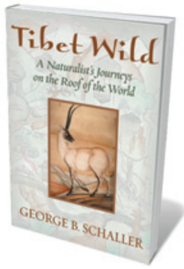
Corporation 2020: Transforming Business for Tomorrow's World
Pavan Sukhdev ISLAND *320 pp. £18.99 (2012)*
Business isn't working — so say a rising number of pundits witnessing cyclical patterns of boom and bust. Green economist and former banker Pavan Sukhdev argues that the corporate model needs an overhaul if profits are to be generated in socially equitable, environmentally benign ways. In his nuanced analysis, corporations need to align their aims with society, becoming viable communities, institutes and financial, human and natural capital 'factories'. His plan for reform focuses on resource taxation, limited leverage, ethical advertising and disclosure of externalities such as pollution.



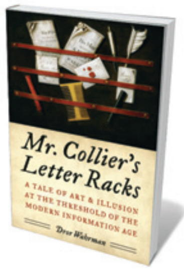
Hyperactive: The Controversial History of ADHD
Matthew Smith REAKTION BOOKS *208 pp. £25 (2012)*
Mild brain damage, sugar, evolutionary hangovers, genes — answers to the question 'What causes ADHD?' are mind-bogglingly diverse. But, argues Matthew Smith in the first medical history of attention-deficit hyperactivity disorder, we may need to accept that explanations will be pluralistic and relativistic. Smith addresses biological, social and psychological issues, from an eighteenth-century description of the fidgets to the first cases, the drugs and the diets. With powerful pharmaceuticals involved and US diagnoses running at 9% a year in 5- to 17-year-olds, this is a timely chronicle.



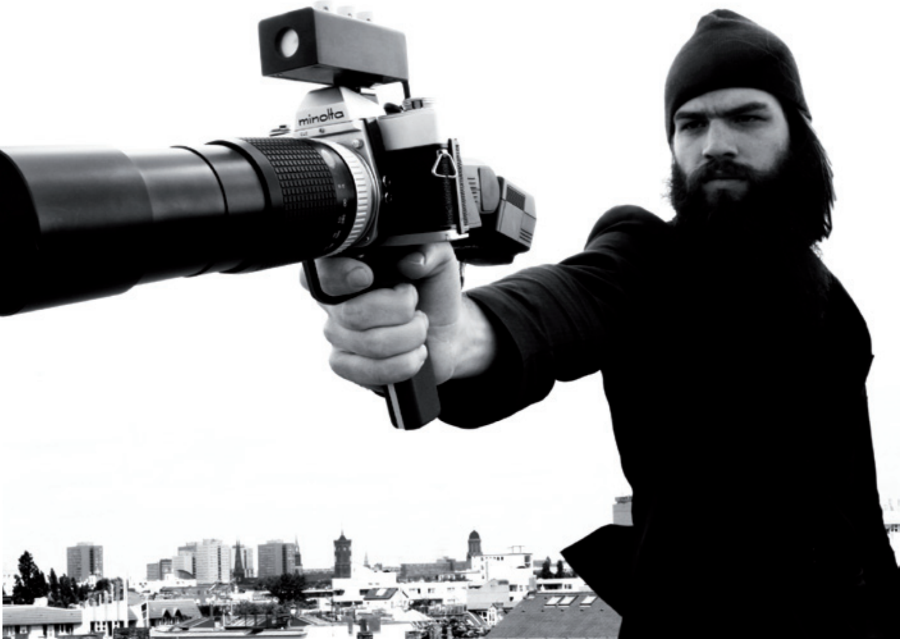
Measurement
Paul Lockhart HARVARD UNIVERSITY PRESS *416 pp. £20 (2012)*
This invitation to tackle mathematical questions is infused with the joys of the rarefied reality of maths. Paul Lockhart largely avoids complex formulae and the wilder shores of jargon, opting instead for simple geometric drawings, lucid instructions and honest warnings about the hurdles. Covering size, shape, space and time, Lockhart, a maths teacher, gets through scores of problems, from showing that a cone in a hemisphere occupies half the volume to determining the size of the largest circle that can sit at the bottom of a parabola. Elegant, amusing and challenging.



Tibet Wild: A Naturalist's Journeys on the Roof of the World
George B. Schaller ISLAND *412 pp. £18.99 (2012)*
After 50 years of research on endangered species, field biologist George Schaller is still swimming against the tide of change in the wild. This highly personal compilation, part memoir and part research record, celebrates that "raw terrain where lakes are the colour of molten turquoise" — the Tibetan Plateau, particularly the northern plain of the Chang Tang. Woven into vivid accounts of tracking mammals such as snow leopards and chiru, or Tibetan antelope, are Schaller's tracings of the impacts of climate change and population growth on one of the last animal strongholds.



Mr. Collier's Letter Racks: A Tale of Art and Illusion at the Threshold of the Modern Information Age
Dror Wahrman OXFORD UNIVERSITY PRESS *352 pp. £22.95 (2012)*
The technology-driven explosion in cheap print 300 years ago spawned the first information age. As historical sleuth Dror Wahrman relates, a little-known Dutchman commented covertly on the metamorphosis — in *trompe l'oeil* paintings. Edward Collier created 'snapshots' of letter racks stuffed with printed speeches and pamphlets — riddled with visual jokes and puzzles that were coded criticisms of the limitations of print and the politics of the day.



Julius von Bismarck with his *Image Fulgurator*, which manipulates other people's photographs.

RICHARD WILHELMER

Q&A Julius von Bismarck
Collision creator

Julius von Bismarck is the first artist in residence at the particle-physics laboratory CERN, near Geneva in Switzerland. As he prepares to give the final lecture of his residency, he talks about whipping mountains, hacking photographs and digging into the history of invention.

How did you get into art?
I was a chaotic kid. I hacked machines, grew marijuana, made bombs and experimented with electronic circuits I found in the garbage. Eventually I had to decide between engineering, physics and robotics, and making useless and abstract works of art. I chose to become an artist because I thought it would allow me the freedom to keep working on both abstract and technical fronts. It turned out to be the right decision. Now I make sculptures and installations, and present performances and many other interventions in public space. I have used a lot of technology, but I'll use any medium I can to put my thoughts into other people's brains.

What did you do as CERN's artist in residence?
Among my projects is a moving sculpture representing the three-dimensional shadow of a rotating hypercube. I was hoping to convey the feeling of trying to imagine something you can't perceive, and to make the extra dimensions required by some scientific theories more accessible. Another involves installing motors in the ceilings of industrial spaces to cause their hanging lamps to oscillate slightly. Together, the lamps can create complex geometric patterns to demonstrate physical ideas, such as the astronomical phenomenon of redshift, which shows that the Universe is expanding.

You created something called the *Perpetual Storytelling Apparatus*. Tell me about that.
German artist Benjamin Maus and I built it to dig into the history of invention. The machine hangs on the wall and sketches patent drawings on a 50-metre-long roll of paper. You can upload a novel and it will translate it, sentence by sentence, into a series of illustrations from the US Patent and Trademark Office's database of 8 million patents. It fills in the gaps between unrelated patents using a six-degrees-of-separation-style algorithm that uses citations to find the shortest path from, say, an atomic power plant to a steam engine. We keep the books we feed into the machine secret, but if you know the book you can follow the story. Science fiction works quite well, because people are constantly talking about strange weapons and new technologies. To illustrate a single book, the apparatus must draw for months on many rolls of paper.

Von Bismarck whips Alpine peaks as part of his *Punishment series 2011* (image cropped).

J. VON BISMARCK

You have also made an *Image Fulgurator*. What is that?
It's a sort of weapon that allows me to manipulate other people's photographs without their knowledge. Triggered by a flash, the device projects an image of my choice onto an object exactly at the moment it is being photographed. As a design student I was thinking about how local authorities can decide where advertisements are permitted, and how I could fight that power by hacking into other people's photographs. The first version looked like a gun, but I've now made it small enough to fit in a normal camera so I can smuggle it into press conferences. I superimposed a dove on the portrait of Mao at Tiananmen Square in Beijing. And I projected a cross onto Obama's podium when he visited Berlin in 2008. When the Pope visited Madrid last summer, I worked with Spanish artist Santiago Sierra to project the word 'NO' above him.

What are you working on now?
I have an ongoing series called *Punishment* in which I film myself punishing the natural world. There is a romantic cliche of nature that has been glorified by artists and advertising agencies alike. To punish nature for this hubris, I have climbed into the Alps and whipped its peaks for hours. In forests I have whipped the trees. On a beach in Brazil I whipped the Atlantic Ocean. It has caused me pain and exhaustion. Rage against nature is a fight you cannot win.

What else did you propose at CERN?
I wanted to make a dent in the surface of Lake Geneva, using a controlled underwater vacuum, to make the viewer think about gravity. We are so used to gravity that we don't perceive it. But without gravity the lake's water would just float around in drops. At CERN I learned that gravity is not yet understood at the microscopic scale; no one has observed a gravity particle yet. A dent in the lake could get people talking about physics.

INTERVIEW BY JASCHA HOFFMAN

Correspondence

Sustainability is key to development goals

As the United Nations General Assembly meets in New York this week, the global community should look beyond the 2015 expiry of the Millennium Development Goals (MDGs). We need to embrace environmental sustainability to alleviate poverty, and to ensure that economic growth does not generate inequality. Development challenges must be addressed worldwide, going beyond the traditional divides of north versus south, or rich versus poor.

The MDGs on global poverty, health, education and gender equality have provided an unprecedented rallying point for action by governments, civil society, international agencies and the private sector. The numbers of people without access to safe drinking water, living in extreme poverty and dying during childbirth have all been halved since 1990. But what comes next is urgent.

We have to tackle the new interlocking realities: inequality is worsening in many areas; 1.3 billion people still live on less than US$1.25 per day; pressure on natural resources is growing; and climate change is upon us. The world is changing profoundly as the middle class expands globally, and more people now live in cities than in rural areas.

New ideas, including the push for Sustainable Development Goals, are emerging that could help to shift the development agenda in the right direction.

Manish Bapna *World Resources Institute, Washington DC, USA.*
mbapna@wri.org

Arab science must help itself

Ehsan Masood urges Arab liberals to help to build Islamic science policy (*Nature* **488,** 131; 2012). Regrettably, academics who participated in the Arab Spring are still constrained by opposing military, religious or even tribal forces.

There are other ways to strengthen scientific progress. These would include increasing national investment in local human resources and boosting support from international stakeholders for developing science and technology.

The Middle East and North Africa must also discard their long history of 'self-loathing', which manifests as deference to Western expertise, undervaluation of distinguished expatriates and resistance to their advice. On my travels around the region, I hear from internationally renowned Arab experts who have tried to offer assistance to local establishments, only to be rebuffed or treated as second class.

The region must convince scientists in the diaspora that it is serious about promoting its own science and education. Most important, it must embrace diversity and freedom of thought.

Mustafa al'Absi *University of Minnesota Medical School, Duluth, Minnesota, USA.*
malabsi@umn.edu

Preprint servers: no author fees

Prepublication of scientific papers on preprint servers such as arXiv.org allows prompt scrutiny of the research by the scientific community (see, for example, go.nature.com/nwjmbt). Because the results are freely accessible, the arXiv approach is infinitely superior to the 'author pays' model of open-access journal publishing that is being pushed as a way to penetrate the paywall.

The number of mathematics and physics papers being posted to arXiv is rapidly increasing; subsequent publication in a journal serves mainly to validate the results.

Making authors pay to publish their research endangers the open and egalitarian nature of the scientific enterprise. Researchers in developing countries, unaffiliated researchers, graduate students and faculty members without large federal grants could all be priced out of publishing their work. The arXiv model offers a sensible and affordable alternative.

Ilya Kapovich *University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.*
kapovich@math.uiuc.edu

Preprint servers: follow arXiv's lead

In the spirited debate on open-access publishing (see, for example, *Nature* **487,** 302; 2012), it is worth remembering that the process is likely to be different in different research fields.

Self-archiving of pre- and postprint versions of research papers in high-energy physics on the arXiv.org server (the 'green' model of open-access publishing) has been running smoothly since 1991. In this model, the peer-review system of the scientific journals serves as a quality stamp. The system works because of institutional subscriptions to journals.

I would encourage scientists in other fields take a closer look at the arXiv model.

Tommy Ohlsson *KTH Royal Institute of Technology, Stockholm, Sweden.*
tohlsson@kth.se

Realizing Australia's bioenergy potential

Andrew Lang and his colleagues present an enthusiastic vision for bioenergy in Australia (*Nature* **488,** 590–591; 2012). However, the country's large area and low population notwithstanding, the contribution of biomass to the national renewable-energy portfolio will be constrained by poor soils, low primary productivity and the time and logistics needed to establish plantings.

Optimistic projections of bioenergy production and usage based on biomass flows and operation costs in other countries can harm the industry by unrealistically raising public expectations. Australia's sustainable biomass production might be enough to replace 15% of electricity demand or 34% of current petroleum consumption (D. R. Farine *et al. GCB Bioenergy* **4,** 148–175; 2012). Investors and decision-makers will need to make real-world tests of assumptions and uncertainties in their own national and local contexts.

Advanced energy-generation technologies based on woody lignocellulose sources are reducing the impact of the bioenergy sector on food crops, but more research, development and investment are needed to find sustainable ways of resolving the competition for land and water.

Realizing Australia's bioenergy potential will take time. Supportive policies are needed to reduce the risk for investment. Regional hotspots for economic and sustainable bioenergy production must be identified and technology locally matched to biomass resources and scale of use. A clear strategy will be needed to incorporate biomass resources effectively into the suite of available renewable energies.

Luis C. Rodriguez, Alexander Herr, Michael H. O'Connor *CSIRO Ecosystem Sciences, Australia.*
luis.rodriguez@csiro.au

CORRECTION

The Outlook article 'Trials of an anticancer jab' (*Nature* **488,** S4–S6; 2012) contained an error in the table: in Australia, the uptake for at least 1 dose is 83% not 74%. The label in the graphic stated that vaccination started in 2010 instead of 2007. And reference 9 should have read Tabrizi, S. N. *et al. J. Infect. Dis.* (in the press).

Neil Armstrong

(1930–2012)

Engineer, pilot, astronaut and the first human to walk on the Moon.

Neil Armstrong accepted that he would always be remembered as the first human to set foot on the Moon, which he did as commander of the *Apollo 11* mission on 20 July 1969. But that was not all that defined him. Armstrong was proud of his naval service: flying combat missions in the Korean War and testing high-performance aircraft. He was a committed educator and a quiet but thoughtful force in delineating US aerospace policy.

Armstrong died aged 82 on 25 August 2012, from complications of heart-bypass surgery. He was born on 5 August 1930 on his grandparents' farm near Wapakoneta, Ohio. His parents took him to air races as a boy and he fell in love with the prospect of flying. Armstrong took his first plane ride in a Ford Tri-Motor at the age of 6, and by 16 he had earned his student pilot's licence; all before he could drive a car or had a high-school diploma.

After high school, Armstrong went to Purdue University in West Lafayette, Indiana, to study aeronautical engineering. His scholarship from the US Navy required him to serve a tour of active duty after two years of education. He became an aviator, and in 1950 was sent to Korea, where he engaged in raids of North Korean railway bridges, targets that have been immortalized in the James Michener novel *The Bridges at Toko-Ri*.

In 1952, Armstrong returned to Purdue to finish his bachelor's degree and then joined the National Advisory Committee for Aeronautics (NACA), which became NASA in 1958. As an engineer and research pilot he worked at NACA's Lewis Research Center near Cleveland, Ohio, and then at the High Speed Flight Research Center in Edwards, California. Armstrong flew pioneering aircraft, including the X-15 rocket plane that set speed and altitude records in the early 1960s. Over the years, he took the controls of more than 200 models of jet, rocket, glider and helicopter.

Armstrong transferred to astronaut status in 1962, and was one of nine members of the second class to be chosen for spaceflight. (The first class, the Mercury Seven, was picked in 1959.) His experience was invaluable during his first mission in March 1966, in *Gemini VIII*, when he and David Scott docked their capsule in orbit around Earth to an Agena spacecraft, the first such rendezvous in space. Soon after, the joined vehicles began tumbling uncontrollably. Armstrong managed to undock *Gemini VIII* and stabilize the craft, and the astronauts made an emergency landing in the Pacific Ocean. They learned later that a stuck control jet had caused the spacecraft to spin.

NASA

On *Apollo 11*, as is now legendary, Armstrong, along with Michael Collins and Edwin 'Buzz' Aldrin, completed the first Moon landing. Armstrong piloted the lunar module during its final descent, and stepped down to make his famous (mis)statement: "That's one small step for [a] man, one giant leap for mankind." Armstrong and Aldrin spent around two and a half hours on the Moon's surface, collecting samples, doing experiments and taking photographs, before returning to Collins and the lunar module. The trio splashed down into the Pacific Ocean on 24 July.

Almost everyone who is old enough recalls where they were when *Apollo 11* touched down on the Moon. In the United States, the landing briefly unified a nation divided by political, social, racial and economic tensions. Millions of people, myself included, imagined being Armstrong as he reached the "magnificent desolation" of the lunar surface. As a 15-year-old, I sat with friends on a car, looking up at the Moon and listening to the astronauts over the car's radio. It was an inspirational moment, but it was fleeting.

What was not fleeting was how Armstrong embodied the spirit of the accomplishment until his last breath. He lived a life of quiet grace, rarely embroiling himself in day-to-day fights while exemplifying a unique merger of the 'Right Stuff' with introspection and calm. Some have characterized him as a recluse; I know some at NASA wished that he had supported the agency's initiatives more publicly.

Armstrong sought neither fame nor riches. When he could have done anything he wished, he chose to teach aerospace engineering at the University of Cincinnati in Ohio.

For four decades, Armstrong made clear his perspective on myriad aerospace issues to many leaders and to the commissions on which he served. His considered opinions carried weight, notably in the Centennial of Flight Commission, which oversaw the commemoration of the Wright brothers, and in the investigation of the *Challenger* accident.

Commentators usually compare Armstrong's first step on the Moon to Christopher Columbus's arrival at the Americas, as vanguards of sustained exploration and settlement. Yet increasingly, the parallel seems to be the voyage centuries earlier of Norse explorer Leif Erikson — a stillborn event in the long process of exploring new lands.

Armstrong was always perplexed by the praise heaped on him. The Moon landing was the result of the labour of hundreds of thousands and the accomplishment of a generation of humanity, he said. We will all miss him, not just for being the first Moon walker, but for the honour and dignity with which he carried the weight of that history on his back. ■

Roger D. Launius *is a senior curator in the Division of Space History at the Smithsonian Institution's National Air and Space Museum. e-mail: launiusr@si.edu*

Intensified Arabian Sea tropical storms

ARISING FROM A. T. Evan, J. P. Kossin, C. E. Chung & V. Ramanathan *Nature* **479,** 94–97 (2011)

Tropical cyclones over the Arabian Sea in the pre-monsoon season (May–June) have intensified since 1997 (ref. 1, Fig. 1a) owing to significant reductions in storm-ambient vertical wind shear (VWS) in the troposphere; these reductions have decreased on average by about 3 m s^{-1} from the pre-1997 epoch (1979–1997) to the recent epoch (1998–2010)[1]. The authors attribute the reduction of pre-monsoon VWS to the dimming effects of increased anthropogenic black carbon and sulphate emissions[1]. However, observations show no sign of a significant decreasing trend in VWS (Fig. 1b), in contrast to the simulated, aerosol-induced decreasing trend in ref. 1. We further show that the decrease of VWS in the recent epoch is caused by substantially advanced (by 15 days) tropical-cyclone occurrences, caused by the early onset of the Asian summer monsoon. There is a Reply to this Brief Communication Arising by Evan, A. T. *et al. Nature* **489**, doi:10.1038/nature11471 (2012).

About 90% of the Arabian Sea pre-monsoon tropical cyclones occur from mid-May to mid-June, during which the mean VWS over the tropical-cyclone intensification zone increases approximately from 12 m s^{-1} to 25 m s^{-1} (Fig. 1c). We discover that the mean date of the lifetime maximum intensity (LMI) of tropical cyclones occurred 15 days earlier from 8 June to 24 May (Fig. 1c) and the mean genesis date of tropical cyclones also shifted earlier by 16 days from 6 June to 21 May. The epochal changes of the mean tropical-cyclone genesis and LMI dates are significant at 99.8% confidence level by Student's *t*-test. Even if we exclude the two earliest storms in the recent epoch, the shift in mean LMI (or genesis) date is 11 (or 13) days, which remains significant at above the 99% confidence level. Using the 15-days-earlier tropical-cyclone LMI date results in a remarkable reduction of the storm-ambient VWS, by about 5.8 m s^{-1} from before 1997 to the recent epoch (Fig. 1c), which is an order of magnitude larger than the model-simulated, aerosol-induced VWS decrease (a difference of about 0.5 m s^{-1} between the two epochs).

What has caused the earlier occurrences of tropical cyclones in the recent epoch? The genesis of tropical cyclones requires favourable environmental conditions[2] and the likelihood of genesis can be estimated by the Genesis Potential Index (GPI)[3]. Computation of the GPI averaged over the main genesis region (8°–20° N, 55°–75° E) reveals that the maximum GPI occurred significantly earlier in the recent epoch (Fig. 1c), which explains why the recent tropical cyclones occurred earlier. Further analysis of the factors controlling the GPI change indicates that the shift of maximum GPI can be attributed mainly to the increased low-level cyclonic shear vorticity, which enhances the boundary-layer moisture convergence and lower tropospheric humidity. Meanwhile, the contribution of ambient VWS to GPI change is moderate. As shown in Fig. 1d, the monthly mean relative vorticity of 850 hPa in May, during which all tropical cyclones were generated in the recent epoch, has significantly increased from before 1997 to the recent epoch. This finding is consistent with the early-onset trend in the Asian summer monsoon[4] and with May (or June) storms being associated with an early (or late) onset of the southwesterly monsoon[5].

We further find that the early development of the southwesterly monsoon may be caused by enhanced land–ocean thermal contrast between the Asian landmass and the equatorial Indian Ocean (Fig. 1d), which can reinforce the northward pressure gradients that in turn strengthen southwesterly monsoon and associated cyclonic shear vorticity.

We consider that the increased land–ocean thermal contrast may be attributed to internally generated interdecadal variation and anthropogenic warming. On the one hand, the upward trends seen in tropical-cyclone intensity, low-level vorticity and the land–ocean thermal contrast all display a significant rapid upswing in the late 1990s (Fig. 1), and these upswings are in phase with the swift phase transition of the Interdecadal Pacific Oscillation[6], suggesting that natural variability could be the cause. On the other hand, previous studies have suggested that (1) increasing greenhouse-gas warming will enhance land–ocean thermal contrast in summer[7] and (2) the

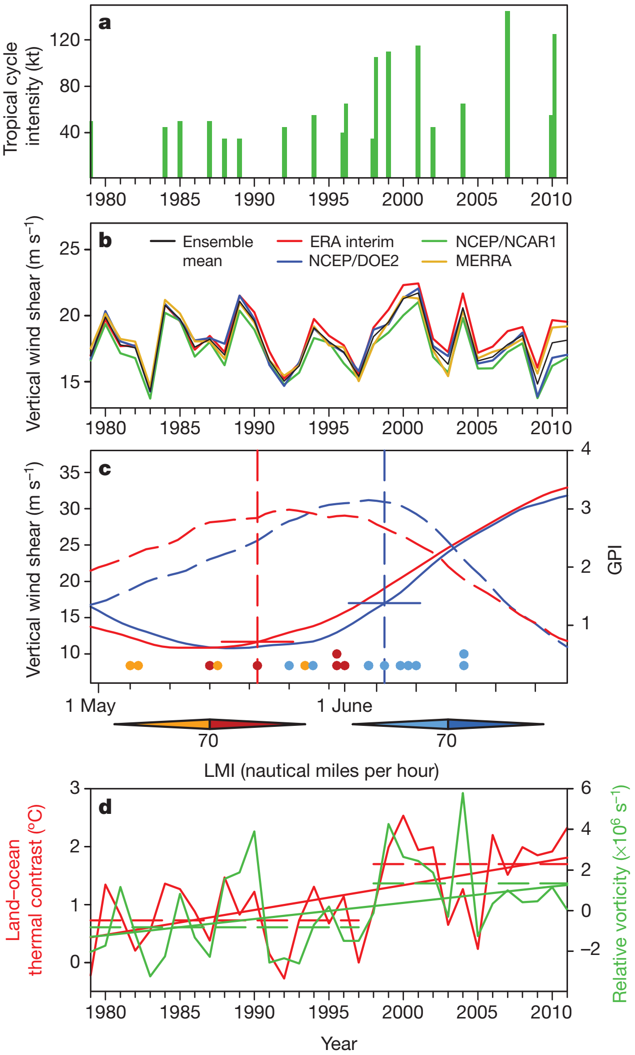


Figure 1 | Cause of the intensified Arabian Sea tropical cyclone. **a**, Intensification of the tropical cyclone shown by the maximum wind speed of each pre-monsoon Arabian Sea tropical cyclones. kt, knot (one nautical mile per hour). **b**, Change of ambient VWS over the Arabian Sea from various reanalysis data sets[11–14] and their ensemble mean. **c**, Epochal changes of tropical-cyclone LMI dates (solid dots), average LMI dates (vertical dashed lines), 15-day running mean VWS (solid curves) and GPI (dashed curves). Red denotes the recent epoch; blue denotes the pre-1997 epoch. Orange/turquoise (or red/blue) shading on the solid dots represents LMI less than (or greater than) 70 nautical miles per hour. **d**, Changes in the 850-hPa relative vorticity in May (green) averaged over the tropical-cyclone genesis region and the land–ocean thermal contrast (red) with a sudden change around 1997 at the 98% and 99% confidence levels, respectively. The thin solid lines denote linear trends and the dashed lines indicate the two epochal means.

increasing loading of desert dust and soot aerosol over northern India and the foothills of the Himalayas during the pre-monsoon season may enhance tropospheric warming and strengthen land–ocean thermal contrast by absorbing solar radiation[8,9]. Further studies are needed to clarify the causes and the complexity of the interaction between aerosols, the monsoon and tropical cyclones.

Methods

Tropical-cyclone data are from the International Best Track Archive for Climate Stewardship (IBTrACS) data[10]. On the basis of the ERA interim[11], NCEP/NCAR1[12], NCEP/DOE2[13], and MERRA[14] reanalysis data sets, the VWS of 850–200 hPa is averaged over the tropical-cyclone intensification zone (10°–22° N, 55°–75° E). The GPI is averaged over the tropical-cyclone genesis region (8°–20° N, 55°–75° E). The land–ocean thermal contrast is defined by the 2-m air temperature difference between the Asian continent (20°–35° N, 40°–80° E) and the equatorial Indian Ocean (10° S–10° N, 40°–80° E). The significance of sudden changes was established using the Le Page test[15].

Bin Wang[1,2], Shibin Xu[1,3] & Liguang Wu[2]
[1]Department of Meteorology and International Pacific Research Center, University of Hawaii at Manoa, Honolulu, Hawaii 96822, USA.
[2]Key Laboratory of Meteorological Disaster of Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, 210044, China.
email: wangbin@hawaii.edu
[3]Physical Oceanography Laboratory, Ocean University of China, Qingdao, 266100, China.

Received 24 April; accepted 13 July 2012.

1. Evan, A. T., Kossin, J. P., Chung, C. E. & Ramanathan, V. Arabian Sea tropical cyclones intensified by emissions of black carbon and other aerosols. *Nature* **479,** 94–97 (2011).
2. Gray, W. M. Global view of the origin of tropical disturbances and storms. *Mon. Weath. Rev.* **96,** 669–700 (1968).
3. Emanuel, K. A. & Nolan, D. S. Tropical cyclones and the global climate system. In *26th Conf. on Hurricanes and Tropical Meteorology* 240–241 (American Meteorological Society, 2004).
4. Kajikawa, Y., Yasunari, T., Yoshida, S. & Fujiyama, H. Advanced Asian summer monsoon onset in recent decades. *Geophys. Res. Lett.* **39,** L03803 (2012).
5. Evan, A. T. & Camargo, S. J. A climatology of Arabian Sea cyclonic storms. *J. Clim.* **24,** 140–158 (2011).
6. Wang, B., Liu, J., Kim, H. J., Webster, P. J. & Yim, S. Y. Recent change of the global monsoon precipitation (1979–2008). *Clim. Dyn.* doi:10.1007/s00382-011-1266-z (2012).
7. Meehl, G. A. *et al.* in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon, S. *et al.*) Ch. 10 (Cambridge University Press, 2007).
8. Lau, K.-M. & Kim, K.-M. Observational relationships between aerosol and Asian monsoon rainfall, and circulation. *Geophys. Res. Lett.* **33,** L21810 (2006).
9. Meehl, G., Arblaster, J. & Collins, W. Effects of black carbon aerosols on the Indian monsoon. *J. Clim.* **21,** 2869–2882 (2008).
10. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS): unifying tropical cyclone data. *Bull. Am. Meteorol. Soc.* **91,** 363–376 (2010).
11. Dee, D. P. *et al.* The ERA-interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137,** 553–597 (2011).
12. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–471 (1996).
13. Kanamitsu, M. *et al.* NCEP-DOE AMIP-II reanalysis (R-2). *Bull. Am. Meteorol. Soc.* **83,** 1631–1643 (2002).
14. Rienecker, M. M. *et al.* MERRA—NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Clim.* **24,** 3624–3648 (2011).
15. Lepage, Y. A table for a combined Wilcoxon Ansari-Bradley statistic. *Biometrika* **60,** 113–116 (1973).

Author Contributions B.W. planned and finalized the manuscript. L.W. drafted the manuscript, S.X. did data analysis and plotting.

Competing Financial Interests Declared none.

doi:10.1038/nature11470

Evan *et al.* reply

REPLYING TO B. Wang, S. Xu & L. Wu *Nature* **489,** http://dx.doi.org/10.1038/nature11470 (2012)

Arabian Sea tropical cyclones have become stronger over the past 30 years owing to a reduction in vertical wind shear (VWS) brought about by radiative forcing from pollution aerosols[1]. Wang *et al.*[2] argue that the decline in VWS results from a systematic shift in storm genesis date, which may be part of a natural cycle or another consequence of regional pollution. However, their conclusions[2], although interesting, are not supported by our analysis and are probably sensitive to biases in the observational record.

Wang *et al.*[2] concluded that there are no linear trends in pre-monsoon VWS observable from reanalysis when averaged over the Arabian Sea. From analyses of three independent reanalysis data sets we showed[1] that there are downward trends in VWS over the northeastern region of the Arabian Sea, where most storms form and track. We did not claim that there existed basin-averaged downward trends in VWS.

Wang *et al.*[2] claim that the strong intensification of tropical cyclones is due to an advancement in their genesis or LMI date. Although the Julian date on which pre-monsoon storms reached their lifetime maximum intensity (LMI) has advanced by 15.6 days, of the five strongest storms forming in the basin ($\mathrm{LMI} > 50\,\mathrm{m\,s^{-1}}$) three had LMI dates greater than the climatological mean (Fig. 1a). There is no statistically significant difference in the mean LMI date (P-value = 0.66) or genesis date (P-value = 0.51) when comparing the strongest storms ($\mathrm{LMI} > 50\,\mathrm{m\,s^{-1}}$) to all others, which is evidence that the weak VWS experienced by the strongest storms is not due to an advancement in genesis or LMI date.

Another flaw in the theory of ref. 2 is their implicit assumption that Arabian Sea tropical cyclones experience the basin-averaged VWS. For example, they[2] used the annual cycle of VWS averaged over the Arabian Sea to show that a systematic change in genesis date is responsible for the decrease in VWS experienced by Arabian Sea storms in the second half of the 1979–2010 record. However, tropical cyclones are influenced only by the VWS in the environment surrounding the storm, or the storm-ambient VWS.

To test the hypothesis of ref. 2 we calculated the storm-ambient VWS from reanalysis data[3], and then for reanalysis VWS fields consisting of only linear trends in VWS (which indicates the influence of

Table 1 | Change in storm-ambient VWS between 1979–1996 and 1997–2010

	Storm-ambient VWS difference		*P* value
	Median	Mean	
Reanalysis	−3.1	−3.2	0.01/0.01
Trends	−3.5	−4.0	0.01/0.02
Genesis	−1.7	−3.4	0.06/0.10

The changes due to trends in the reanalysis VWS (Trends), and due to changes in the genesis date (Genesis) in the median/mean pre-monsoon Arabian Sea tropical-cyclone storm-ambient VWS from the first (1979–1996) to the second (1997–2010) half of the tropical cyclone record in the reanalysis (AMIP[3]) are shown. The *P*-values indicate the statistical significance of the separation of the median/mean as determined by a Wilcoxon rank sum/*t*-means test.

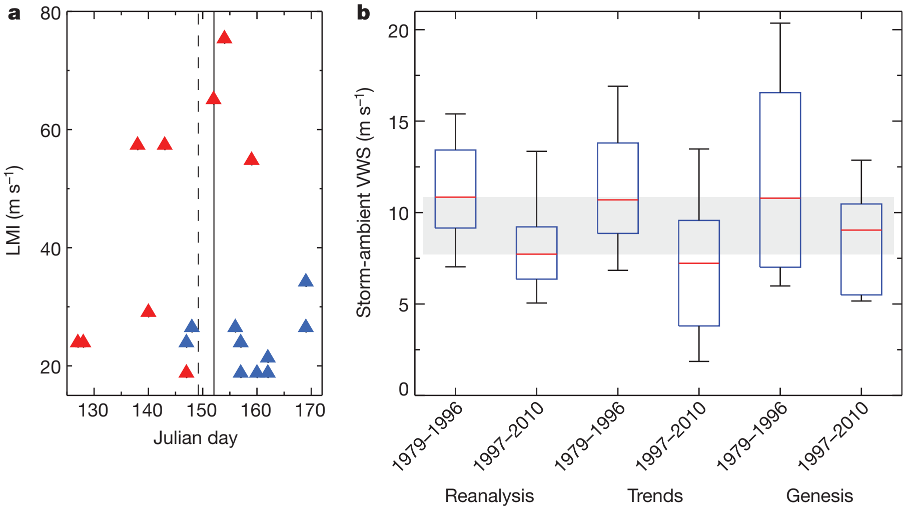


Figure 1 | Arabian Sea cyclone LMI dates and influence of trends and genesis date on storm-ambient VWS. a, The Arabian Sea pre-monsoon cyclone LMI is plotted as a function of the Julian date on which LMI occurred. Blue triangles indicate storms in 1979–1997; red triangles indicate storms in 1998–2010. The solid vertical line is the mean LMI date for the storms with LMI < 50 m s^{-1}; the dashed vertical line is the mean LMI date for the storms with LMI > 50 m s^{-1}. **b**, Box plots of storm-ambient VWS from reanalysis VWS (Reanalysis), from only the trends in reanalysis VWS (Trends), and from only the annual cycle in reanalysis VWS (Genesis) are shown. Plotted are the median (red line), the inner quartile range (blue box), and the 25th (or 75th) percentile minus (or plus) 1.5 times the inner quartile range (horizontal black line), with the grey shaded region indicating the change in storm-ambient VWS from the Reanalysis case for reference. The significance (*P*-value) of the separation of the median and mean values in each distribution is given in Table 1. To facilitate comparison we added an offset value to the storm-ambient VWS data from the trend and annual cycle cases (in **b**) so that the median storm-ambient VWS in each was equivalent to that from the Reanalysis case.

VWS trends on changes in storm-ambient VWS) or only the annual cycle of VWS (which indicates the influence of tropical-cyclone genesis date on changes in storm-ambient VWS) (see Methods). A comparison of the periods 1979–1996 and 1997–2010 shows that the median storm-ambient VWS changed by −3.1 m s^{-1} (Fig. 1b)[1]. We find a change of −3.5 m s^{-1} in the median storm-ambient VWS from linear trends in the reanalysis VWS, and a change of −1.9 m s^{-1} from earlier tropical cyclone genesis dates. Therefore, the background linear trends in VWS are large enough to account for the noted change in storm-ambient VWS, but the change in storm genesis date can account for only about half of the observed reduction in storm-ambient VWS. Furthermore, there is no statistically significant (*P*-value < 0.05) change in the mean or median of the distribution of storm-ambient VWS owing to earlier genesis date (Table 1).

We further note that the tendency for earlier genesis over the past 30 years is probably due in part to technological improvements. Before 1998 the Arabian Sea was observed at the edge of the viewing area of the existing geostationary satellites, but in 1998 the repositioning of a satellite over the Indian Ocean improved the angle at which storms in the basin could be viewed. This improvement in the observations probably increased the ability to detect and track weak storms[4], contributing to the increase in tropical cyclogenesis date shown in ref. 2.

Finally, there are a number of studies confirming the influence of pollution aerosols on the characteristics of the Southeast Asia monsoon[5–8], supporting our findings that these pollution aerosols are potentially the fundamental cause of the uptick in Arabian Sea cyclone intensity[1].

Methods

Tropical cyclogenesis is defined as the time at which a tropical storm acquired a sustained wind speed of 17 m s^{-1}, and LMI date is the time at which the storm first reached its maximum sustained wind speed, based on data from the International Best Track Archive for Climate Stewardship[9]. We calculated the storm-ambient VWS from reanalysis[3] in a manner consistent with ref. 1. The influence of linear trends (or genesis date) on the storm-ambient VWS was estimated by removing the annual cycle and interannual variability (or interannual variability and linear trends) from the reanalysis data, retaining only the linear trends (or annual cycle) in the pre-monsoon VWS fields. In the genesis case, and as in ref. 2, we calculated the storm-ambient VWS for tropical cyclones forming before (or on or after) 1997 using the annual cycle of VWS calculated from 1979–1996 (or 1997–2010).

Amato T. Evan[1,2], James P. Kossin[3,4], Chul 'Eddy' Chung[5] & V. Ramanathan[2]
[1]University of Virginia, Charlottesville, Virginia 22904, USA.
email: aevan@ucsd.edu
[2]Scripps Institution of Oceanography, San Diego, California 92093, USA.
[3]NOAA's National Climatic Data Center, Asheville, North Carolina 28801, USA.
[4]NOAA Cooperative Institute for Meteorological Satellite Studies, Madison, Wisconsin 53706, USA.
[5]Gwangju Institute of Science and Technology, Gwangju 500712, South Korea.

1. Evan, A. T., Kossin, J. P., Chung, C. E. & Ramanathan, V. Arabian Sea tropical cyclones intensified by emissions of black carbon and other aerosols. *Nature* **479,** 94–97 (2011).
2. Wang, B., Xu, S. & Wu, L. Intensified Arabian Sea tropical storms. *Nature* **489,** http://dx.doi.org/10.1038/nature11470 (2012).
3. Kanamitsu, M. *et al.* NCEP-DOE AMIP-II Reanalysis (R-2). *Bull. Am. Meteorol. Soc.* **83,** 1631–1643 (2002).
4. Evan, A. T. & Camargo, S. J. A climatology of Arabian Sea cyclonic storms. *J. Clim.* **24,** 140–158 (2011).
5. Chung, C. E. & Ramanathan, V. Weakening of North Indian SST gradients and the monsoon rainfall in India and the Sahel. *J. Clim.* **19,** 2036–2045 (2006).
6. Lau, K.-M. *et al.* The joint aerosol-monsoon experiment. *Bull. Am. Meteorol. Soc.* **89,** 369–383 (2008).
7. Meehl, G., Arblaster, J. & Collins, W. Effects of black carbon aerosols on the Indian monsoon. *J. Clim.* **21,** 2869–2882 (2008).
8. Ramanathan, V. *et al.* Atmospheric brown clouds: impacts on South Asian climate and hydrological cycle. *Proc. Natl Acad. Sci. USA* **102,** 5326–5333 (2005).
9. Knapp, K. P., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS): unifying tropical cyclone data. *Bull. Am. Meteorol. Soc.* **91,** 363–376 (2010).

doi:10.1038/nature11471

ASTRONOMY

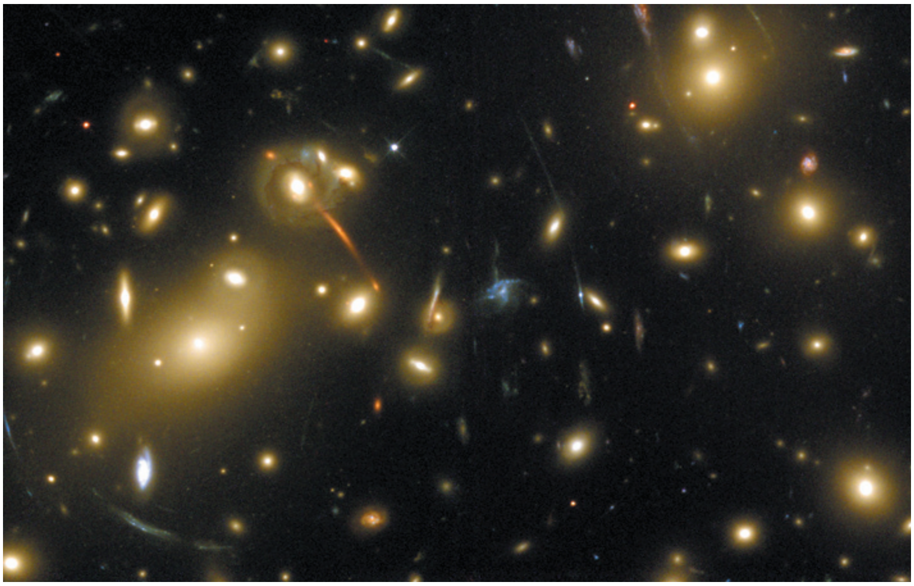
Searching for the cosmic dawn

The Hubble Space Telescope, teaming up with a 'cosmic lens', has revealed a highly magnified galaxy thought to date back to 500 million years after the Big Bang. The find provides a glimpse of the first stages of galaxy formation. SEE LETTER P.406

DANIEL STARK

Astronomers have long sought to trace the history of the Universe from its origins to the present day. Our earliest picture comes from the study of the cosmic microwave background radiation, which provides a portrait of the Universe when it was less than 400,000 years old. With not a single star yet formed at that time, the Universe was shrouded in darkness and permeated by newly created hydrogen atoms. The next available detailed picture comes nearly one billion years later and reveals a dramatically different landscape. Not only were galaxies containing billions of stars common, but the hydrogen that once filled most of space had become highly ionized between these galactic systems. In this issue, Zheng and colleagues[1] (page 406) help to fill in the intervening time with the discovery of a galaxy that pushes the cosmic frontier back to just 500 million years after the Big Bang. When the galaxy's photons departed nearly 13.2 billion years ago, the Universe was less than 4% of its current age. By studying this early galaxy, Zheng *et al.* offer insight into when and how the first galaxies assembled, and whether the energetic radiation they produce was responsible for the 'reionization' of intergalactic hydrogen.

Zheng and colleagues' discovery follows in the footsteps of an exciting period that was ushered in by the installation in 2009 of Wide Field Camera 3 on the Hubble Space Telescope, which provided astronomers with a phenomenal leap forward in infrared imaging capability. Because the expansion of the Universe stretches the wavelength of light by a factor of $1+z$, where z is the redshift of an object such as a galaxy, observations in the infrared regime are a key domain in which to discover the earliest galaxies. To identify such systems, astronomers make use of the fact that light with wavelengths shorter than the redshifted hydrogen Lyman-α line limit, which is $0.1216\,(1+z)$ micrometres, is absorbed by intervening hydrogen clouds. For galaxies within the first 650 million years of cosmic history (redshifts greater than 8), such Lyman-α absorption extinguishes all light at optical wavelengths. So, by searching for galaxies that have a 'break' in their flux between the optical and near-infrared parts of the electromagnetic spectrum, astronomers can identify those galaxies that are most likely to lie at great distances.



A. FRUCHTER (STSCI) *ET AL.*/WFPC2/HST/NASA

Figure 1 | A gravitational lens. Clusters of galaxies can act as cosmic lenses, bending and magnifying the light from distant background sources. This image by the Hubble Space Telescope of the massive galaxy cluster Abell 2218 reveals many lensed galaxies, one of which is so distant that its light left it when the Universe was just 800 million years old. Zheng *et al.*[1] studied another galaxy cluster lens, MACS 1149+2223 (see Fig. 2 of the paper), which has revealed a distant galaxy thought to date back to 500 million years after the Big Bang.

But even with the deepest images yet obtained[2,3] by Hubble's infrared camera, it has proved extremely difficult to break through to the first 500 million years of cosmic time. By identifying the characteristic Lyman-α absorption described above, researchers have unveiled[3] more than 100 galaxies thought to lie between 650 million and 850 million years after the Big Bang, but only one galaxy had been found that could be dated back to 500 million years[4].

To combat the difficulties imposed by the faintness of distant galaxies, Zheng *et al.* used a clever phenomenon called gravitational lensing. This technique relies on the principle that light rays from distant galaxies are bent and often magnified as they pass through the vicinity of massive objects on their way to Earth. By pointing telescopes towards such massive cosmic lenses (Fig. 1), for example a cluster of nearby galaxies, it is possible to detect distant galaxies that are bright enough for detailed study owing to the boost provided by gravitational lensing[5,6].

Zheng and colleagues have been using Hubble's infrared camera to systematically search for distant magnified galaxies behind some of the most massive nearby galaxy clusters. After analysing 12 galaxy clusters using the Lyman-α absorption technique, their efforts finally yielded success, uncovering a galaxy thought to lie just 500 million years after the Big Bang. The foreground cluster of galaxies magnifies the galaxy's light by a factor of 15, allowing its properties to be dissected in greater detail than if it had been found by conventional methods.

Crucial to the authors' finding were observations conducted with the Spitzer Space Telescope, which probes infrared light from old stars. These observations indicate that the galaxy had a significant component of old stars. The authors estimate that stars had been forming in the galaxy for up to 200 million years, building up a stellar mass 150 million times that of the Sun. If this system is representative of galaxies at this redshift, it would suggest that vigorous star formation was already occurring in galaxies by 300 million to 500 million years after the Big Bang. The energetic radiation emitted by these systems could ionize a significant fraction of intergalactic hydrogen within just 500 million years of the Big Bang, consistent with expectations from measurements of the polarization of the cosmic microwave background radiation[7].

Zheng and colleagues' discovery will stimulate further searches for galaxies at this early epoch, and much work remains to be done. Currently, the number of sources that can be dated back to 500 million years after the Big Bang (just two[1,4]) is too small for reliable measures of their number density to be extracted. Moreover, without spectroscopic observations to complement the images, the galaxies' distances from Earth cannot be determined unambiguously. Some progress on both fronts is expected in the coming years from surveys conducted with the Hubble and Spitzer telescopes, as well as with new infrared spectrographs that have been installed on ground-based telescopes.

Within the next decade, however, the exploration of galaxies in the early Universe will be transformed by the construction of giant ground-based telescopes with apertures of 20–40 metres and by the launch of the James Webb Space Telescope. These powerful facilities will not only dramatically increase the number of galaxies known throughout the first 500 million years, but will also provide the spectroscopic capability necessary to confirm their distances. Through spectroscopy of highly magnified galaxies such as that reported by Zheng *et al.*, these studies can even begin to reveal the galaxies' chemical make-up and the kinematics of the gas they contain, leading to a much-improved understanding of when galaxies emerged and how their radiation contributed to the reionization of hydrogen. ■

Daniel Stark *is in the Department of Astronomy, University of Arizona, Tucson, Arizona 85721, USA.*
e-mail: dpstark@email.arizona.edu

1. Zheng, W. *Nature* **489,** 406–408 (2012).
2. Robertson, B. E., Ellis, R. S., Dunlop, J. S., McLure, R. J. & Stark, D. P. *Nature* **468,** 49–55 (2010).
3. Bouwens, R. J. *et al. Astrophys. J.* **737,** 90 (2011).
4. Bouwens, R. J. *et al. Nature* **469,** 504–507 (2011).
5. Kneib, J.-P., Ellis, R. S., Santos, M. R. & Richard, J. *Astrophys. J.* **607,** 697–703 (2004).
6. Bradley, L. D. *et al. Astrophys. J.* **747,** 3 (2012).
7. Komatsu, E. *et al. Astrophys. J. Suppl. Ser.* **192,** 18 (2011).

NEUROSCIENCE

Attention is more than meets the eye

Our brains focus on important events and filter out distracting ones. An investigation in monkeys reveals a surprising dissociation between the neuronal and behavioural manifestations of attention. SEE LETTER P.434

ALEXANDRA SMOLYANSKAYA
& RICHARD T. BORN

Why are drivers more likely to have an accident if they are talking on a mobile phone? The obvious answer is that they are not paying attention to the road. But what is attention, and what goes on in our brains when we are 'paying' it? For decades, psychologists have proposed that we can direct something rather like a mental spotlight towards particular regions of our surroundings, and that this selectively enhances our perceptual sensitivity in that region. For example, if you were instructed to attend to an area to your left (without looking there), you would be able to detect a dimmer light in that region than elsewhere. Indeed, neurophysiologists have shown that attention amplifies the responses of neurons whose preferred spatial region (the neuron's 'receptive field') corresponds to the attended region[1]. However, on page 434 of this issue, Zénon and Krauzlis[2] report that, in monkeys, inactivating a brain structure called the superior colliculus impairs visual attention but retains the enhanced responses of neurons in the brain's cerebral cortex.

In addition to amplifying — or, more precisely, increasing the gain of — neurons' responses, attention tends to make the relevant neurons slightly less 'noisy' and more independent (of their neighbours) in their responses; both changes allow them to collectively encode sensory information more reliably. All of these changes make sense, and seem to account for why an animal's perception is enhanced by attention[3]. This is why Zénon and Krauzlis's results[2] will come as a surprise to many. By inactivating the superior colliculus, which has previously been shown to be important for attention[4], they impaired the monkeys' ability to detect a visual target while ignoring an irrelevant, distracting stimulus in another part of the visual field.

Inactivation of the superior colliculus (by injecting a drug that inhibited neuronal activity) did not create a basic sensory deficit, like a blind spot, because the impairment was evident only when there were competing stimuli[5]. But, despite the severe impairment in the animals' ability to pay attention to the relevant stimulus, all of the known neural correlates of attention (including increased gain) in two sensory areas in the cerebral cortex — the middle temporal area (MT) and the medial superior temporal area (MST) — remained intact. Thus, the authors uncoupled the neuronal changes that are thought to underlie attention from the behavioural manifestation of attention.

What are we to make of this? For a start, we can conclude that the superior colliculus is not the only source of signals responsible for the changes in early sensory areas (those closer to the sensory receptors). We can also conclude that the improvements in the encoding of sensory information in the MT and MST are not sufficient to produce the perceptual effects of attention.

However, although Zénon and Krauzlis found no changes in any of the known neural correlates of attention, it is conceivable that they missed the right neurons — we know that, at any given location in a sensory area, only a subset of neurons contributes to any given task[6]. Moreover, it is possible that inactivation of the superior colliculus impairs the attentional system in other ways, and that the neuronal changes in the MT and MST are insufficient to overcome the deficit. For example, if selective attention emerges as the result of competition among visual representations in multiple brain regions[7], the increases in gain in the MT and MST might simply be overbalanced by the loss of enhancements in other areas more directly connected to the superior colliculus.

Another possibility is that attention follows a two-stage mechanism: a first stage produces the gain changes in early sensory areas, whereas a later stage selects among these enhanced signals. In this model, the superior colliculus would act as part of the selection filter, the activity of which determines whether signals from a particular sensory region will be used to guide behaviour. Without the superior colliculus, the corresponding part of visual space is effectively filtered out, or ignored, as it is for patients with brain damage who have 'unilateral neglect'[8] — they may fail to eat food from one side of their plate, for example, or to shave one side of their face.

Such a filtering stage would explain why humans often miss large changes in the visual scene. In one famous example, observers are asked to count the number of passes of a basketball among teammates, and they fail to notice a person in a gorilla suit who wanders

50 Years Ago

Concern is expressed about continuing traffic in heroin. In some areas, limitation of the use of opium appears to have encouraged opium addicts to turn to heroin, which has been more readily available. The controls on the illicit production of heroin and on the traffic of this drug need to be enforced more strictly. Much stress is laid by the [World Health Organization Expert] Committee on the necessity for providing the medical profession as early as possible with complete and accurate information on the addiction-producing and habit-forming properties of new drugs and on their therapeutic properties. The further development of media for disseminating such information should be encouraged.
From *Nature* 22 September 1962

100 Years Ago

Man is worth many sparrows; he is worth all the animal population of the globe, and if there were not room for both, the animals must go. I will pass no judgment on those who find the keenest pleasure of life in gratifying the primeval instinct of sport. I will admit that there is no better destiny for the lovely plumes of a rare bird than to enhance the beauty of a beautiful woman … But I do not admit the right of the present generation to careless indifference or to wanton destruction. Each generation is the guardian of the existing resources of the world; it has come to a great inheritance, but only as a trustee … [T]here is no resurrection or recovery of an extinct species, and it is not merely that here and there one species out of many is threatened, but that whole genera, families, and orders are in danger.
From *Nature* 19 September 1912

through the scene[9]. This happens even though neurons in an early visual area accurately represent — and can therefore be used to detect — the gorilla and all such highly salient changes. At some stage, even these otherwise obvious events are filtered out, presumably to focus processing on the behaviourally relevant information. We speculate that inactivation of the superior colliculus, as described by Zénon and Krauzlis, may impair this latter stage.

It must be that a brain area other than the superior colliculus is responsible for the gain changes observed in MT and MST neurons. One possible candidate is a region of the cortex known as the frontal eye fields, which are involved in visual attention and eye movements. Indeed, there is evidence that electrical stimulation of the frontal eye fields can produce gain changes in early sensory areas similar to those produced by attention[10]. Future experiments will be necessary to determine how the activities of the superior colliculus and those of areas such as the frontal eye fields are coordinated to converge on an attended location. In particular, as the convergence of enhanced signals has been proposed to occur in a region of the parietal cortex called the lateral intraparietal area[11], it will be important to determine whether inactivation of the superior colliculus leads to more-pronounced deficits in the effects of attention on neurons in this area than those observed in the MT and MST.

Zénon and Krauzlis's results suggest that there are at least two cooperating stages: attentional-gain modulation and subsequent selection. Their work therefore calls for further studies of how such systems interact to endow us with a mechanism that we depend on every day: the option to ignore our mobile phones and focus on the road ahead. ■

Alexandra Smolyanskaya *and* **Richard T. Born** *are in the Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA.*
e-mail: richard_born@hms.harvard.edu

1. Reynolds, J. H. & Chelazzi, L. *Annu. Rev. Neurosci.* **27,** 611–647 (2004).
2. Zénon, A. & Krauzlis, R. J. *Nature* **489,** 434–437 (2012).
3. Cohen, M. R. & Maunsell, J. H. R. *Nature Neurosci.* **12,** 1594–1600 (2009).
4. Kustov, A. A. & Robinson, D. L. *Nature* **384,** 74–77 (1996).
5. Lovejoy, L. P. & Krauzlis, R. J. *Nature Neurosci.* **13,** 261–266 (2010).
6. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. *Vis. Neurosci.* **13,** 87–100 (1996).
7. Desimone, R. & Duncan, J. *Annu. Rev. Neurosci.* **18,** 193–222 (1995).
8. Brain, W. R. *Brain* **64,** 244–272 (1941).
9. www.youtube.com/watch?v=IGQmdoK_ZfY
10. Moore, T. & Armstrong, K. M. *Nature* **421,** 370–373 (2003).
11. Bisley, J. W. & Goldberg, M. E. *Annu. Rev. Neurosci.* **33,** 1–21 (2010).

MATERIALS CHEMISTRY

Liposomes derived from molecular vases

Liposomes are ubiquitous components of skin moisturizers and other personal-care products. Modified liposomes prepared from receptor-like molecules open up fresh opportunities for therapeutic and industrial applications.

CYRUS R. SAFINYA & KAI K. EWERT

The imaginations of diverse groups of scientists, from physicists to pharmacologists, have been captured by liposomes — simple mimics of highly complex cell membranes. Typical liposomes are spheres with walls consisting of bilayers of amphiphilic lipids (molecules that have hydrophilic, polar head groups and hydrophobic, non-polar tails). Their unique structure enables them to trap hydrophobic molecules within their bilayer and hydrophilic molecules within their interior (Fig. 1a). Writing in *Chemical Communications*, Kubitschke *et al.*[1] add another dimension to this cargo-carrying ability with their report of liposomes derived from vase-shaped cavitands[2], which are receptor-like molecules that wrap around 'guest' compounds. The cavitands can encapsulate these guest molecules and present them at high densities at the liposome surface, a capability that might be useful for drug delivery.

Liposomes — also known as vesicles — were serendipitously discovered in 1964 during investigations of phospholipids[3]. The demonstration of their encapsulation properties, and the remarkable structural resemblance between liposomes and cell membranes in electron micrographs, led to the realization that lipids form the main permeability barriers of biological membranes. Today, by far the largest use of liposomes and their encapsulating properties is in the multibillion-dollar personal-care industry[4], as moisturizers and carriers of nutrients in gels and cream formulations. But they have also emerged as a research tool, within which biologists can isolate and molecules inside the cavitand are from outside the liposome.

Further development of cavitand liposomes as drug-delivery vehicles will undoubtedly see the addition of stealth and cell-targeting properties. In fact, Kubitschke *et al*. have already used eight short chains of poly(ethylene glycol) — the water-soluble polymer that forms the repulsive shell of most stealth liposomes — to line the rim of their cavitands so that the molecules retain their binding properties in water (Fig. 1d).

A class of vesicle related to liposomes is the polymersome[15] — vesicles that are made from amphiphilic polymers, rather than lipids. Another possible extension of the authors' work would therefore be the development of polymersomes formed from cavitands that have hydrophilic and hydrophobic polymers attached at opposite ends. Polymersomes are tougher than liposomes, and can sustain greater deformation before rupture. Aside from chemical delivery, cavitand polymersomes would therefore be suitable for applications in which assemblies of judiciously chosen guest molecules undergo large rates of deformation, such as molecular coatings that have controlled friction properties. ■

Cyrus R. Safinya *and* **Kai K. Ewert** *are in the Departments of Materials, Physics, and Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, Santa Barbara, California 93106, USA. They are also affiliated with the Materials Research Laboratory, University of California, Santa Barbara.*
e-mails: safinya@mrl.ucsb.edu; ewert@mrl.ucsb.edu.

1. Kubitschke, J., Javor, S. & Rebek, J. *Chem. Commun.* **48,** 9251–9253 (2012).
2. Cram, D. J. *Science* **219,** 1177–1183 (1983).
3. Bangham, A. D. & Horne, R. W. *J. Mol. Biol.* **8,** 660–668 (1964).
4. Lasic, D. D. *Liposomes: From Physics to Applications* (Elsevier, 1993).
5. Papahadjopoulos, D. in *Stealth Liposomes* (eds Lasic, D. D. & Martin, F.) 1–6 (CRC, 1995).
6. Felgner, P. L. *et al. Proc. Natl Acad. Sci. USA* **84,** 7413–7417 (1987).
7. Ewert, K. K. *et al. Top. Curr. Chem.* **296,** 191–226 (2010).
8. Nabel, G. J. *et al. Proc. Natl Acad. Sci. USA* **90,** 11307–11311 (1993).
9. www.wiley.com/legacy/wileychi/genmed/clinical
10. Huang, L., Hung, M.-C. & Wagner, E. (eds) *Non-Viral Vectors for Gene Therapy* 2nd edn, Part I (Academic, 2005).
11. Safinya, C. R., Ewert, K. K. & Leal, C. *Liq. Cryst.* **38,** 1715–1723 (2011).
12. Rädler, J. O., Koltover, I., Salditt, T. & Safinya, C. R. *Science* **275,** 810–814 (1997).
13. Liu, Y., Liao, P., Cheng, Q. & Hooley, R. J. *J. Am. Chem. Soc.* **132,** 10383–10390 (2010).
14. Feher, K. M., Hoang, H. & Schramm, M. P. *N. J. Chem.* **36,** 874–876 (2012).
15. Discher, B. M. *et al. Science* **284,** 1143–1146 (1999).

HUMAN BEHAVIOUR

A cooperative instinct

Acting on a gut feeling can sometimes lead to poor decisions. But it will usually support the common good, according to a study showing that human intuition favours cooperative, rather than selfish, behaviour. SEE LETTER P. 427

SIMON GÄCHTER

In a recent bestselling book, psychologist and Nobel laureate Daniel Kahneman presents a wealth of evidence that much of human decision-making is governed by fast and automatic intuitions, rather than by slow, effortful thinking[1]. Intuitions can sometimes lead us astray, such as when it comes to processing statistical information, but our 'gut feelings' also serve us well in many common situations. One interesting question to ask is how intuition influences social decisions that pit self-interest against collective benefit. Does intuition support cooperation, or do people need time to reflect before deciding to pull their weight? On page 427 of this issue, Rand *et al.*[2] present evidence that the intuitive human reaction is to cooperate, whereas reasoning makes people somewhat more selfish.

If evolution favours self-interest, then people should be equipped with intuitions that help them figure out how to maximize their individual gain[3]. However, recent research in the behavioural sciences challenges the idea that people are mostly selfish[4]. Some theories to explain variation in individual behaviour, based on social preferences[5], assume that people differ in their motivation to act in a cooperative manner[6], but not in their reasoning style. Furthermore, psychological studies have suggested that moral judgements are often made intuitively[7], and because many people view 'freeloading' on other people's contributions as morally blameworthy[8], it is plausible that moral intuitions support cooperation.

To investigate directly the role of intuitions in cooperation, Rand and colleagues used a series of ten public-goods game experiments. In these games, people can choose to either keep an allocation of resources for themselves, or contribute all or a portion of their allocation to a collective pool, which is then distributed evenly among all players. The authors conducted some of the games using an international group of subjects sourced from an online labour market (Amazon Mechanical Turk)[9], and others were conducted in person in the laboratory.

Because intuitions are quickly available, whereas deliberation takes time, Rand *et al.* started by investigating the link between response time and contributions. Previous research on response time across a variety of decisions shows that people choose intuitive options more quickly than those requiring cognitive effort[10], and results from a simple sharing experiment suggest that faster choices are more selfish[11]. However, this is not what Rand and colleagues found in their online experiments. Instead, their results indicate that contributions and decision time are negatively correlated — the faster half of the decision-makers contributed, on average, about 67% of their allocated resources, whereas the slower half contributed about 53%. The authors also detected a similar relationship between response time and cooperation in experiments conducted in person, so the observed correlation seems to be robust.

But correlations are of course no proof of causation. To try to plausibly demonstrate a causal link, Rand and colleagues put the game players under time pressure and observed how this affected their decisions. Previous results from bargaining-game experiments suggest that time pressure leads to fairer outcomes[12] and also increases the likelihood that a person will impulsively reject an unfair offer[13,14]. Furthermore, having to decide under time pressure is stressful, and stress also increases pro-social behaviour[15]. So it is clear that time pressure, which favours intuitions over reflection, influences social considerations. Rand *et al.* show that this extends to cooperation: in their experiments, people under time pressure contributed significantly more than those who made their decisions with no time limit or with a forced delay. Thus, it seems that forcing a person to decide more rapidly — by intuition — increases their tendency to cooperate.

In a final set of experiments, the authors used a writing task to prime participants to think intuitively or reflectively before performing the public-goods game. They found that those primed to use intuition contributed more than those put in reflective mode. Rand and colleagues also found that people who consider their interaction partners in daily life to be cooperative cooperate more when primed to use intuition than when primed to use reflection. This result is consistent with a point made by economics Nobel laureate Herbert Simon, who said that "intuition is nothing more and nothing less than recognition"[16]. Thus, it seems that when people are accustomed to cooperative partners, they develop cooperative intuitions.

Rand and colleagues' study raises interesting concepts for experiments in the social sciences, both in terms of questions that would be worthy of further investigation and how to conduct such experiments. For example, their findings suggest that the common practice of asking participants comprehension questions before an experiment will provide conservative estimates of cooperativeness, because the questioning will put people into reflective mode, which Rand and colleagues have shown is likely to result in them behaving less cooperatively. So is this questioning practice justified? It may be in many cases, such as in studies of people's economic decisions, as economists are typically interested in reflected behaviour.

The study also indicates that intuitions may be particularly important in novel situations, and that experience might trigger reflection that either supports or modifies the initial intuitions. Should economic theories based on social motivations[5] take intuitions into account even if the main importance of intuition is (only) in initiating cooperation? Future research may clarify this question. Furthermore, the authors observe that many (but not all) people are cooperative whether deciding quickly or slowly, intuitively or reflectively, and time pressed or not. For example, even in the experiments in which Rand *et al.* recorded the biggest difference between intuitive and reflective contributions, the contributions made under reflective conditions exceeded the difference added by intuition. Economic and evolutionary theories should attempt to explain these findings.

Finally, existing research suggests that some people are selfish free-rider types, whereas others are conditional cooperators who are willing to contribute if others do so[6]. This observation needs to be squared with Rand and colleagues' results: might it be that conditional cooperators are intuitively cooperative and selfish people take a reflected free ride? The authors have demonstrated that, on average, our intuition is to cooperate, but further studies are needed to understand the variation in this behaviour between individuals. ■

Simon Gächter *is in the Centre for Decision Research and Experimental Economics, University of Nottingham, Nottingham NG7 2RD, UK.*
e-mail: simon.gaechter@nottingham.ac.uk

1. Kahneman, D. *Thinking, Fast and Slow* (Allen Lane, 2011).
2. Rand, D. G., Greene, J. D. & Nowak, M. A. *Nature* **489,** 427–430 (2012).
3. Moore, D. & Loewenstein, G. *Social Justice Res.* **17,** 189–202 (2004).
4. Bowles, S. & Gintis, H. *A Cooperative Species: Human Reciprocity and its Evolution* (Princeton Univ. Press, 2011).
5. Fehr, E. & Schmidt, K. M. in *Handbook of the Economics of Giving, Altruism and Reciprocity* (eds Kolm, S.-C. & Ythier, J. M.) Ch. 8 (Elsevier, 2006).
6. Fischbacher, U., Gächter, S. & Quercia, S. *J. Econ. Psychol.* **33,** 897–913 (2012).
7. Haidt, J. *The Righteous Mind. Why Good People are Divided by Politics and Religion* (Allen Lane, 2012).
8. Cubitt, R. P., Drouvelis, M., Gächter, S. & Kabalin, R. *J. Public Econ.* **95,** 253–264 (2011).
9. Horton, J. J., Rand, D. G. & Zeckhauser, R. J. *Exp. Econ.* **14,** 399–425 (2011).
10. Rubinstein, A. *Econ. J.* **117,** 1243–1259 (2007).
11. Piovesan, M. & Wengström, E. *Econ. Lett.* **105,** 193–196 (2009).
12. Cappelletti, D., Güth, W. & Ploner, M. *J. Econ. Psychol.* **32,** 940–950 (2011).
13. Grimm, V. & Mengel, F. *Econ. Lett.* **111,** 113–115 (2011).
14. Sutter, M., Kocher, M. & Strauß, S. *Econ. Lett.* **81,** 341–347 (2003).
15. von Dawans, B., Fischbacher, U., Kirschbaum, C., Fehr, E. & Heinrichs, M. *Psychol. Sci.* **23,** 651 (2012).
16. Simon, H. A. *Psychol. Sci.* **3,** 150–161 (1992).

MATERIALS SCIENCE

The matryoshka effect

By tailoring the architecture of a bulk material at several different length scales, the ability of a semiconductor to convert heat into voltage has been optimized to a groundbreaking level of performance. See Letter p.414

TOM NILGES

The story of thermoelectric materials, which convert heat into electric voltage, began in the early 1950s as part of the scientific plans for the first manned mission to the Moon. An effective, simple and long-lasting energy source was needed to supply the astronauts at their destination. The solution — a thermoelectric generator based on lead telluride — is still working today on the Moon's Mare Tranquillitatis. On page 414 of this issue, Biswas *et al.*[1] describe what is probably the ultimate optimization of the thermoelectric properties of lead telluride, 43 years after that historic Moon landing. They have doubled the efficiency of the material compared with that used in the first generator, a feat that is not only a tremendous step for one group, but also a giant leap for thermoelectrics.

Thermoelectric generators consist of several 'stacks' — devices in which multiple semiconductor blocks are sandwiched between two electrodes. Each stack produces an electric potential difference if there is a stable, long-lasting temperature difference across it. Two types of semiconductor are needed: an n-type semiconductor, in which a material is 'doped' with a small amount of another material to produce an excess of electrons; and a p-type semiconductor, in which doping produces an excess of positively charged voids called holes, which can act as charge carriers.

The semiconductor blocks are arranged so that opposite sides are connected to different electrodes. If a thermoelectric stack is heated on one side, a potential difference is created by the transfer of electrons (or holes) within the device from the hot to the cold end. In this set-up, the device converts thermal energy into electric energy. Alternatively, if a current is supplied to such a device, then the electric energy can be used to generate a temperature difference between the two sides. In other words, the stack acts as a cooling device.

The improvement of existing thermoelectric materials to achieve more effective energy conversion, or the development of new ones, is a demanding task for chemists, materials scientists and engineers. In general, the thermoelectric process within a material and its efficiency are related to three properties: the Seebeck coefficient, which defines the material's ability to generate a potential difference in response to a temperature difference; the electrical conductivity, a measure of the transport

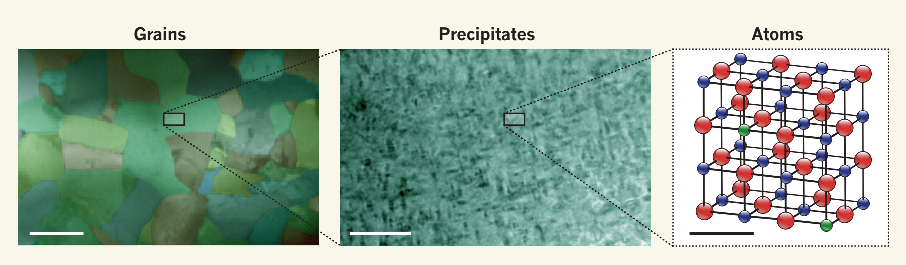


Figure 1 | Better by design. Biswas *et al.*[1] have optimized the thermoelectric properties of lead telluride by controlling its structure at many different length scales. For best performance, the material must contain: grains at the mesoscale (hundreds to thousands of nanometres); nanoscale precipitates of an additive, strontium telluride (several tenths to a few nanometres); and trace amounts of sodium (green atoms), inserted into the material's lattice of lead (blue) and tellurium (red) atoms. The approach works by reducing the thermal conductivity of the material. Scale bars (left to right): 1,000, 50 and 0.5 nanometres.

of electrons or holes through the material; and the thermal conductivity, which defines how well the material transports or equilibrates heat. Semiconductors that have a reasonably large specific electrical conductivity (in the range of thousands of siemens per centimetre) and a passable Seebeck coefficient (hundreds of microvolts per kelvin) are ideal candidates for efficient thermoelectric power generation, but only if the thermal conductivity is small enough to retain the necessary temperature difference effectively.

Different strategies have been developed to optimize these properties. Doping has commonly been used to increase the concentration of mobile charge carriers and holes, or to manipulate the electronic structure of semiconductors. This strategy has worked well in the case of lead telluride, leading to the development of heavily doped substances such as $PbTe_{1-x}Se_x$ (ref. 2; Se is selenium), and to a class of thermoelectric[3] materials known as lead antimony silver tellurides. The Seebeck coefficients of thermoelectric materials can also be improved by tailoring their electronic structures[4].

Most efforts to improve the thermoelectric properties of materials, however, have involved the reduction of thermal conductivity. This requires sophisticated methods, such as altering the nanometre-scale structure of a bulk material (in some cases generating well-defined low-dimensional substructures, such as quantum dots and quantum wells), or forming precipitates of another substance within a thermoelectric material. These approaches prevent heat transport[5] by scattering phonons — the heat carriers in thermoelectric materials. But at least some of these widely used scattering procedures will be hard to scale up for the manufacture of commercial products in the near future.

The brilliance of Biswas and co-workers' study of lead telluride is that they have canalized almost every known strategy for optimizing thermoelectric materials into one system. They used a fast, highly effective technique known as spark plasma sintering (SPS) to synthesize bulk lead telluride, identified strontium telluride as the most suitable candidate to form nanoscale precipitates during the synthesis, and determined the optimal amount of sodium to use as a dopant. This combined approach improves the thermoelectric performance of lead telluride to previously unattainable levels.

The success of the authors' strategy depends on the interplay and occurrence of units at several different length scales: from mesoscopic grains of lead telluride at the micrometre scale, to nanoscale precipitates of strontium telluride, and all the way down to dopants that act at the atomic scale (Fig. 1). The authors call this interplay a panoscopic approach, but the embedding of progressively smaller subunits within the material reminds me of matryoshka (Russian) dolls.

Materials scientists have long dreamt of a fast, reliable method for producing bulk thermoelectrics that does not require complicated optimization procedures and intensive material structuring. Biswas and colleagues' work certainly provides a practical method for making bulk lead telluride, but it also shows us that we really do need to screen for the best additives and dopants, and to tailor structural units, to realize the panoscopic approach for every thermoelectric system. It should also be noted that lead telluride is toxic — for commercial applications, other thermoelectric materials must be found that are non-toxic and inexpensive.

Nevertheless, I am sure that the authors' findings will trigger exponential progress in the performance of thermoelectric materials in general. Indeed, I believe that many surprising and encouraging aspects of thermoelectric behaviour will be discovered as a result of their work. If so, then thermoelectric devices might eventually be improved until their efficiency becomes at least comparable to that of other state-of-the-art energy-conversion devices, such as those that convert solar or geothermal energy. ■

Tom Nilges *is in the Department of Chemistry, Technische Universität München, 85747 Garching, Germany.*
e-mail: tom.nilges@lrz.tum.de

1. Biswas, K. *et al. Nature* **489,** 414–418 (2012).
2. Pei, Y. *et al. Nature* **473,** 66–69 (2011).
3. Hsu, K. F. *et al. Science* **303,** 818–821 (2004).
4. Nilges, T. *et al. Nature Mater.* **8,** 101–108 (2009).
5. He, J. *et al. J. Am. Chem. Soc.* **132,** 8669–8675 (2010).

EVOLUTIONARY BIOLOGY

Insects converge on resistance

In a remarkable example of convergent evolution, insect species spanning 300 million years of divergence have evolved identical single-amino-acid substitutions that confer resistance to plant cardenolide toxins.

NOAH K. WHITEMAN & KAILEN A. MOONEY

Plants and herbivorous insects are the most diverse groups of multicellular organisms, and understanding this species profusion is a central problem in evolutionary biology. A key explanatory hypothesis[1] is that iterative co-evolution between plants that produce toxic compounds[2] (such as nicotine and mustard oils) and herbivorous insects that have resistance to these toxins[3] drives the diversification of each group. Some toxins have an extremely broad mode of action, leading to the evolution of highly divergent detoxification mechanisms in the specialized herbivores that resist these compounds[4]. But other toxins have just one molecular target. The medically important cardiac glycosides (cardenolides), for example, block activity of animal $(Na^+{+}K^+)$ATPase[5], an enzyme that regulates ion gradients across the cell membrane. Writing in *Proceedings of the National Academy of Sciences*, Dobler *et al.*[6] demonstrate how co-evolution between plants containing toxic cardenolides and the herbivorous insects that feed on them represents an exquisite case of convergent molecular evolution, in which distantly related insect species have evolved a common adaptive response in a single gene.

Cardenolides comprise a diverse group of triterpenoid-derived steroids and are produced by at least 60 genera from 12 families of flowering plants, including *Asclepias* species, or milkweeds, and *Digitalis* species, or foxgloves. They disrupt the function of the $(Na^+{+}K^+)$ATPase by binding to the first extracellular loop of the enzyme's α-subunit. Dobler and colleagues studied the sequence of the gene encoding the $(Na^+{+}K^+)$ATPase in 18 cardenolide-resistant herbivorous insect species from 15 genera in 4 orders (Coleoptera, Lepidoptera, Diptera and Hemiptera). They found sequence changes leading to amino-acid substitutions in the α-subunit's first extracellular loop in 16 of the cardenolide-feeding species. The authors then used *in vitro* assays to show that cells expressing some of these altered sequences survive in the presence of the cardenolide oubain, which is otherwise toxic to cultured cells.

Remarkably, one of the amino-acid substitutions (at position 122) is present in cardenolide-feeding species of all 4 orders, but not in any of 14 insect species (also from all 4 orders) that do not feed on cardenolide-producing plants. This suggests that resistance-conferring substitutions present in a subset of cardenolide-feeding species are the result of adaptive evolution that occurred repeatedly across 300 million years of insect evolutionary divergence. However, it is worth noting that some cardenolide-resistant species do not show adaptive changes in the

(Na$^+$+K$^+$)ATPase sequence, suggesting that there are probably other molecular routes to this common trait.

In addition to being a striking case of convergent molecular evolution, this adaptation to cardenolide-producing plants carries ecological implications that extend beyond plant–herbivore interactions (Fig. 1). Some cardenolide-resistant insects sequester these compounds, which provides protection against predators, and some also display warning coloration that advertises this defence[7]. Indeed, the majority of the cardenolide-feeding insects studied by Dobler *et al*. exhibit warning coloration and are highly toxic. These herbivore traits in turn have their own evolutionary impacts. Coexisting toxic species often evolve similar warning signals, presumably to spread the costs of 'teaching' predators avoidance — an evolutionary process described as Müllerian mimicry. And, in a process known as Batesian mimicry, palatable species sometimes evolve to mimic toxic ones, in order to usurp predator protection.

Although these phenomena — host plant specialization, sequestration of plant toxins, warning coloration, and mimicry — are quite common across herbivorous insects, cardenolide-feeding species have been central to the study of these dynamics. The work by Dobler *et al.* therefore adds a crucial molecular-genetic piece to this complex puzzle, establishing cardenolide-feeding herbivores as one of the best examples of how molecular, functional and ecological convergence can be linked.

However, fascinating questions remain. It is unclear whether the evolutionarily important mutations are fixed in each cardenolide-feeding species or whether intraspecies genetic variation exists. Comparing the ratio of non-synonymous changes (mutations that lead to an amino-acid substitution) and synonymous changes (mutations that do not change the amino acid) in the gene encoding the (Na$^+$+K$^+$)ATPase within and between species would help to reveal whether natural selection has fixed many other mutations — in addition to the functionally important ones already identified — in the cardenolide-feeding lineages[8]. Reconstructing the progression of all amino-acid substitutions in each species lineage could reveal whether interactions between mutations have played a part in the adaptive evolution occurring at this sequence[9]. Furthermore, there is variation in the abundance and structure of the cardenolides produced by plants, but we do not know whether this variation has been driven by population variation in insect resistance or vice versa.

It also remains to be confirmed that the amino-acid substitutions identified by Dobler and colleagues do actually confer resistance to cardenolides in insects, because the authors tested the effects of these changes only *in vitro* in human cells. One approach to move beyond cell lines would be to engineer *Drosophila* fruitflies to express the resistance genes, and then conduct oubain-feeding trials. Solving the crystal structure of just one (Na$^+$+K$^+$)ATPase from a cardenolide specialist would also allow for a more refined analysis of the molecular mechanisms underlying each substitution that confers resistance.

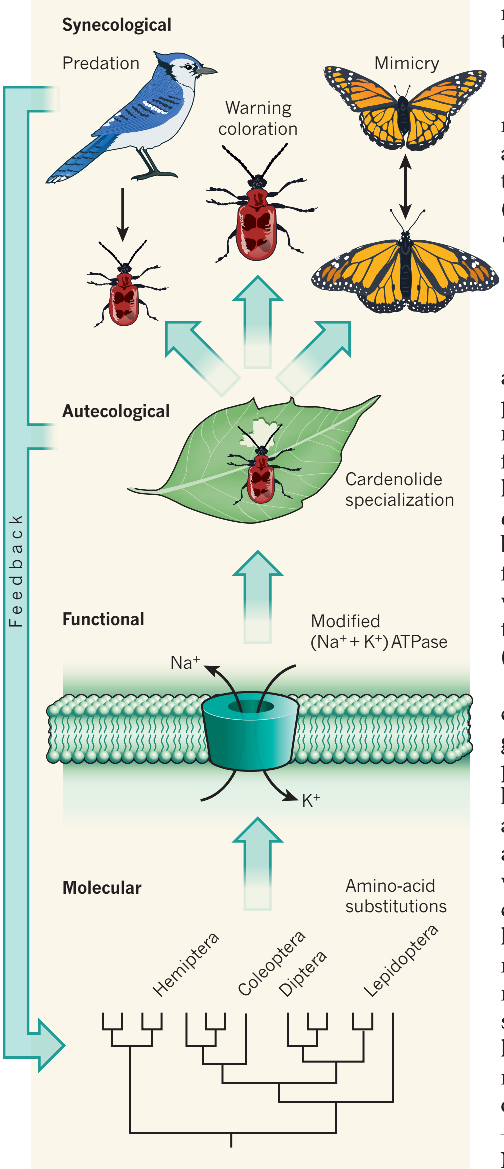


Figure 1 | Hierarchy of adaptation. Dobler *et al.*[6] have identified a single amino-acid substitution in the gene encoding the enzyme (Na$^+$+K$^+$)ATPase in insect species belonging to four orders, which span 300 million years of evolutionary divergence. This (and other) substitutions cause modifications to the enzyme that confer resistance to toxic compounds called cardenolides, possibly allowing insects with this adaptation to feed on plants producing these toxins. This is an example of molecular convergence (the same genetic change provides the same functional outcome in multiple species) leading to autecological convergence (multiple species displaying the same ecological trait; in this case, the ability to feed on cardenolide-producing plants). Such processes can also have synecological effects — those that influence the ecology of other species. For example, some insect species sequester these same plant toxins to protect themselves from predator attack, and display warning coloration to advertise this. Coexisting insects can then evolve to mimic one another, adopting similar warning signals. These ecological effects can, in turn, feed back to influence molecular and functional convergence.

Finally, the findings of Dobler *et al.* pose additional ecological questions. For example, have similar convergent evolutionary mechanisms evolved in the predators that feed on cardenolide-laden insects? Black-headed grosbeak birds (*Pheucticus melanocephalus*) predate overwintering monarch butterflies (*Danaus plexippus*) in the Oyamel fir forests of Mexico[10], but it is not known whether these birds, and other such predators, have also evolved cardenolide-resistant (Na$^+$+K$^+$)ATPases.

The patterns identified by Dobler and colleagues in the insect (Na$^+$+K$^+$)ATPase gene are likely to become a textbook example of convergent evolution at the molecular level — illuminating how a common selective agent can lead to a common set of evolutionary solutions. Their approach highlights the way in which modern evolutionary biology can pinpoint variations in genotype that lead to variation in molecular and organismal traits in ecologically relevant species, not just in model organisms[11]. Ultimately, such studies are helping to identify the ecological forces and molecular mechanisms responsible for generating the diversity of life on Earth. ■

Noah K. Whiteman *is in the Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA.*
Kailen A. Mooney *is in the Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, California 92697, USA.*
e-mails: whiteman@email.arizona.edu; mooneyk@uci.edu

1. Ehrlich, P. R. & Raven, P. H. *Evolution* **18,** 586–603 (1965).
2. Fraenkel, G. *Science* **129,** 1466–1470 (1959).
3. Dethier, V. G. *Evolution* **8,** 33–54 (1954).
4. Winde, I. & Wittstock, U. *Phytochemistry* **72,** 1566–1575 (2011).
5. Rasmann, S. & Agrawal, A. A. *Ecol. Lett.* **14,** 476–483 (2011).
6. Dobler, S., Dalla, S., Wagschal, V. & Agrawal, A. A. *Proc. Natl Acad. Sci. USA* **109,** 13040–13045 (2012).
7. Bernays, E. A. *BioScience* **48,** 35–44 (1998).
8. McDonald, J. H. & Kreitman, M. *Nature* **351,** 652–654 (1991).
9. Aardema, M. L., Zhen, Y. & Andolfatto, P. *Mol. Ecol.* **21,** 340–349 (2012).
10. Fink, L. S. & Brower, L. P. *Nature* **291,** 67–70 (1981).
11. Dean, A. M. & Thornton, J. W. *Nature Rev. Genet.* **8,** 675–688 (2007).

ARTICLE

doi:10.1038/nature11406

Bose glass and Mott glass of quasiparticles in a doped quantum magnet

Rong Yu[1], Liang Yin[2], Neil S. Sullivan[2], J. S. Xia[2], Chao Huan[2], Armando Paduan-Filho[3], Nei F. Oliveira Jr[3], Stephan Haas[4], Alexander Steppke[5], Corneliu F. Miclea[6,7], Franziska Weickert[6], Roman Movshovich[6], Eun-Deok Mun[6], Brian L. Scott[6], Vivien S. Zapf[6] & Tommaso Roscilde[8]

The low-temperature states of bosonic fluids exhibit fundamental quantum effects at the macroscopic scale: the best-known examples are Bose–Einstein condensation and superfluidity, which have been tested experimentally in a variety of different systems. When bosons interact, disorder can destroy condensation, leading to a 'Bose glass'. This phase has been very elusive in experiments owing to the absence of any broken symmetry and to the simultaneous absence of a finite energy gap in the spectrum. Here we report the observation of a Bose glass of field-induced magnetic quasiparticles in a doped quantum magnet (bromine-doped dichloro-tetrakis-thiourea-nickel, DTN). The physics of DTN in a magnetic field is equivalent to that of a lattice gas of bosons in the grand canonical ensemble; bromine doping introduces disorder into the hopping and interaction strength of the bosons, leading to their localization into a Bose glass down to zero field, where it becomes an incompressible Mott glass. The transition from the Bose glass (corresponding to a gapless spin liquid) to the Bose–Einstein condensate (corresponding to a magnetically ordered phase) is marked by a universal exponent that governs the scaling of the critical temperature with the applied field, in excellent agreement with theoretical predictions. Our study represents a quantitative experimental account of the universal features of disordered bosons in the grand canonical ensemble.

Disorder can have a very strong effect on quantum fluids. Owing to their wave-like nature, quantum particles are subject to interference when scattering against disordered potentials. This leads to their quantum localization (or Anderson localization), which prevents—for example—electrons from conducting electrical currents in strongly disordered metals[1], and non-interacting bosons from condensing into a zero-momentum state[2]. Yet interacting bosons represent a matter wave with arbitrarily strong nonlinearity, whose localization properties in a random environment cannot be deduced from the standard theory of Anderson localization. It has been predicted[3,4] that for strongly interacting bosons, Anderson localization manifests itself in the Bose glass: in this phase, the collective modes of the system—and not the individual particles—are Anderson-localized over arbitrarily large regions, leading to a gapless energy spectrum, and a finite compressibility of the fluid. Moreover, nonlinear bosonic matter waves should undergo a localization–delocalization quantum phase transition in any spatial dimension when the interaction strength is varied[3,4]; the transition brings the system from a non-interacting Anderson insulator to an interacting superfluid condensate, or from a superfluid to a Bose glass. Such a transition is relevant for a large variety of physical systems, including superfluid helium in porous media[5], Cooper pairs in disordered superconductors[6,7], and cold atoms in random optical potentials[2,8]. Despite the long history of activity on the subject, a quantitative understanding of the phase diagram of disordered and interacting bosons based on experiments is still lacking.

Recent experiments have demonstrated the capability of realizing and controlling novel Bose fluids made of quasiparticles in condensed matter systems (ref. 9 and 10, and references therein). In this context, a prominent place is occupied by the equilibrium Bose fluid realized in quantum magnets subjected to a magnetic field (ref. 10, and references therein) in which disorder can be introduced in a controlled way by chemical doping, leading to novel bosonic phases[11–15]. The ground state of such systems without disorder and in zero field corresponds to a gapped bosonic Mott insulator. Extra bosons can be injected into the system by applying a critical magnetic field that overcomes the gap, and that drives a transition to a superfluid state (a magnetic Bose–Einstein condensate, BEC). Such a state corresponds to an XY antiferromagnetic state of the spin components transverse to the field. Here we investigate the Bose fluid of magnetic quasiparticles realized in the model compound $NiCl_2{\cdot}4SC(NH_2)_2$ (DTN)[16] with spin $S = 1$ via experiments (a.c. magnetic susceptibility, d.c. magnetization and specific heat), and large-scale quantum Monte Carlo (QMC) simulations. Disorder is introduced by Cl → Br substitution, which, as we will see, leads to randomness in the bosonic hoppings and interactions. We select this compound because the parent compound (pure DTN) has been shown to exhibit Bose–Einstein condensation of the spin system with high accuracy[17]. We also select it because it can be doped very cleanly, which is extremely unusual among similar quantum magnets. The Cl atom sits in an over-sized cage such that it can be replaced by a larger Br atom with very minimal changes in the lattice constants and no observable structural distortion (see Supplementary Information). Thus we can use Br substitution to modify bosonic parameters (for example, magnetic exchange and crystalline electric fields) without other unwanted effects, such as local changes in site symmetry and local modulations of the lattice constant. In experiments and QMC simulations, we observe a Bose glass in two extended regions of the temperature–magnetic field phase diagram of Br-doped DTN. The gapless nature of the Bose glass manifests itself in a finite uniform magnetic susceptibility (corresponding to the

[1]Department of Physics and Astronomy, Rice University, Houston, Texas 77005, USA. [2]Department of Physics and National High Magnetic Field Laboratory, University of Florida, Gainesville, Florida 32611, USA. [3]Instituto de Fisica, Universidade de São Paulo, 05315-970 São Paulo, Brazil. [4]Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90089-0484, USA. [5]Max Planck Institute for Chemical Physics of Solids, Nöthnitzer Strasse 40, 01187 Dresden, Germany. [6]Condensed Matter and Magnet Science, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [7]National Institute for Materials Physics, 077125 Bucharest-Magurele, Romania. [8]Laboratoire de Physique, Ecole Normale Supérieure de Lyon, CNRS UMR5672, 46 Allée d'Italie, 69364 Lyon, France.

compressibility of the quasiparticles), and in a non-exponential decay of the specific heat at low temperature, probing the low-energy density of states. This gapless state extends down to zero field: in this limit the compressibility/susceptibility vanishes while the spectrum remains gapless, giving rise to a Mott glass[18–21]. We investigate the thermodynamic signatures of the Mott and Bose glasses, and the Bose-glass-to-superfluid transition, which is characterized by a novel universal exponent for the scaling of the condensation temperature with applied field.

Magnetic properties of pure DTN

The magnetic properties of pure DTN are those of antiferromagnetic $S = 1$ chains of Ni^{2+} ions, oriented along the crystallographic c axis, and coupled transversely in the a–b plane[16,22]. (The structure of DTN is actually that of two interpenetrating tetragonal lattices, which can be considered effectively as decoupled[23]). The magnetic Hamiltonian is given by

$$\begin{aligned}\mathcal{H} = J_c \sum_{\langle ij \rangle_c} \mathbf{S}_i \cdot \mathbf{S}_j + J_{ab} \sum_{\langle lm \rangle_{ab}} \mathbf{S}_l \cdot \mathbf{S}_m \\ + D \sum_i \left(S_i^z\right)^2 - g\mu_B H \sum_i S_i^z\end{aligned} \qquad (1)$$

where $\mathbf{S}_i = (S_i^x, S_i^y, S_i^z)$ is the spin operator at site i, $J_c = 2.2$ K is the antiferromagnetic coupling for bonds $\langle ij \rangle_c$ along the c axis, $J_{ab} = 0.18$ K is the coupling for bonds $\langle lm \rangle_{ab}$ in the a–b plane, and $D = 8.9$ K is the single-ion anisotropy. μ_B is the Bohr magneton, g is the gyromagnetic factor along the c axis, and H is the applied magnetic field. Here we use a value $g = 2.31$ which is larger by 2% with respect to the value quoted in ref. 22. This value allows us to obtain the best agreement between the experimental and theoretical magnetization curves. In zero field, the large D forces the system into a quantum paramagnetic state with each spin close to its $|m_S = 0\rangle$ state (m_S being the S^z eigenvalue). Mapping the $S = 1$ spin states onto bosonic states with occupation $n = m_S + 1$, the quantum paramagnet corresponds to a Mott insulator of bosons with $n = 1$ particles per Ni site, and with a gap $\Delta \approx D - 2J_c - 4J_{ab} + \mathcal{O}(J_c^2/D)$ for the addition of an extra boson. A magnetic field exceeding the value $H_{c1}^{(0)} = \Delta / g\mu_B \approx 2.1$ T is able to close the spin gap and to create a finite density of excess bosons that condense into a magnetic BEC (see Fig. 1a). The appearance of excess bosons translates into a finite magnetization along the field axis; their long-range phase coherence translates into long-range XY antiferromagnetic order transverse to the field. Long-range order persists up to a critical condensation temperature T_c which, for $H \gtrsim H_{c1}^{(0)}$, scales with the applied field as $T_c \propto \left|H - H_{c1}^{(0)}\right|^{\phi}$. Here $\phi = 2/3$, as predicted by mean-field theory for a diluted gas of excess bosons, and as measured with very high accuracy down to 1 mK (ref. 17). When the magnetic field is increased further, the spins are brought to saturation for $H = H_{c2}^{(0)} = (D + 4J_c + 8J_{ab})/g\mu_B = 12.3$ T, and the system transitions from a BEC to a Mott insulator with $n = 2$ particles per site. Correspondingly, the BEC critical temperature vanishes as $T_c \propto \left|H - H_{c2}^{(0)}\right|^{\phi}$.

Experimental phase diagram of Br-doped DTN

We have measured the critical temperatures and fields for magnetic Bose–Einstein condensation in $Ni(Cl_{1-x}Br_x)_2 \cdot 4SC(NH_2)_2$ (referred to as Br-DTN) with $x = 0.08 \pm 0.005$ by measuring the a.c. susceptibility at low frequencies and the specific heat (see Supplementary Information). Measurements of a.c. and d.c. susceptibility were performed at fixed temperature and varying fields, and they show a step-like increase/decrease corresponding to the critical field for Bose–Einstein condensation, similar to the pure sample[17,24] (see Fig. 2a and b). The main difference compared to pure DTN is that—at low temperatures—the upper and lower edge of the steps are rounded by disorder; as we will see below, this rounding is a fundamental indication of the nature of the phases connected by the transition. An independent estimate of the critical Bose–Einstein condensation temperature as a function of the field is obtained by the location of a sharp λ-peak in the specific heat (Fig. 2c). The remarkable sharpness of the features in the specific heat corresponding to the BEC transition supports the fact that true long-range order persists despite the strong doping introduced in the system. Moreover, for temperatures below the λ-peak the specific heat clearly follows a T^3 behaviour, consistent with long-range XY antiferromagnetic order in three dimensions.

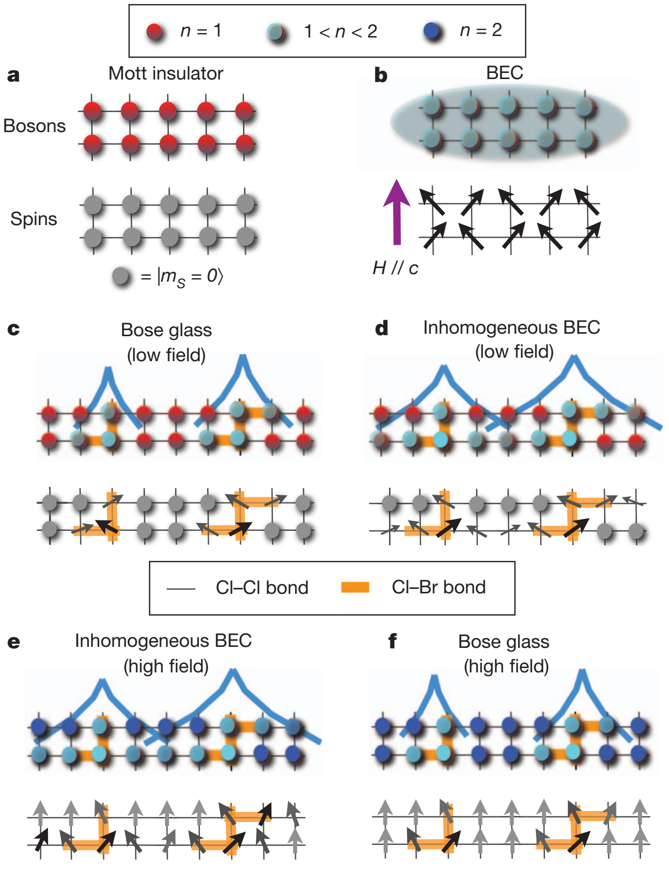


Figure 1 | Sketch of the bosonic phases of DTN and Br-doped DTN. In the undoped case, an increasing magnetic field along the c axis (purple arrow) drives the system from a Mott insulating phase (**a**) to a BEC phase (**b**) by injecting delocalized excess bosons (indicated in cyan) on top of the Mott insulating background at density $n = 1$. In the doped case, an arbitrarily weak magnetic field can inject extra bosons in the rare Br-rich regions (indicated by the orange bonds) which are localized and incoherent in the (low-field) Bose glass phase (**c**)—their localized wavefunction is sketched by the light-blue lines, and the corresponding local orientations of the spins are sketched by the arrows (the darker the arrow, the larger the fluctuating transverse moment induced by the field). Further increasing the magnetic field leads to the percolation of phase coherence via coherent tunnelling of the excess bosons between the localized regions, giving rise to an inhomogeneous BEC (**d**). For strong magnetic fields $H \lesssim H_{c2}$ the spins away from the Br-bonds are close to saturation/double occupancy (represented in dark blue), and unpolarized spins/singly occupied sites, corresponding to bosonic holes, only survive in the Br-rich regions (**e**). These holes are localized into disconnected, mutually incoherent states when entering the high-field Bose glass (**f**).

Figure 3 summarizes the experimental phase diagram of Br-DTN. Br doping has a profound affect on the phase diagram of DTN: in particular both the lower and upper critical fields for the onset of magnetic Bose–Einstein condensation at $T \to 0$ are found to shift to lower values, $H_{c1} = 1.07(1)$ T and $H_{c2} = 12.16(1)$ T, as shown in Fig. 3. But most importantly the magnetic behaviour of Br-DTN

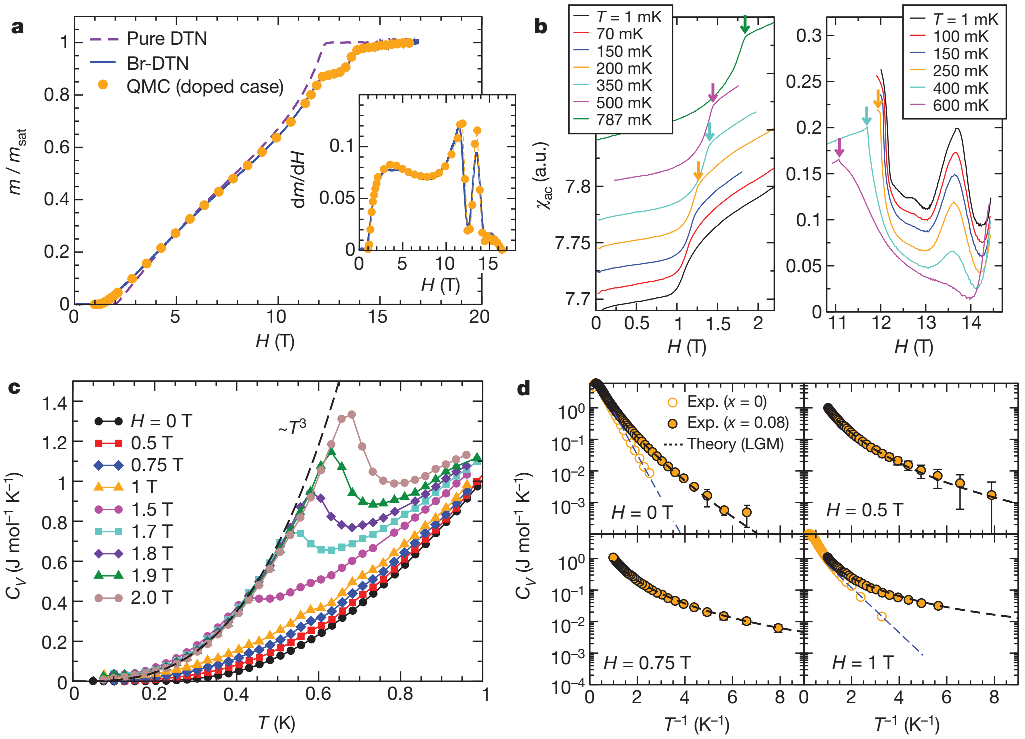


Figure 2 | Thermodynamic properties of the magnetic Bose glass and BEC phases. a, Magnetization curve of Br-DTN at $T = 19$ mK, compared to QMC results, and to pure DTN magnetization (measured at $T = 16$ mK). Inset, the d.c. susceptibility curve, obtained by differentiating the magnetization. **b**, a.c. susceptibility of Br-DTN at frequency $f = 88.7$ Hz close to the lower and upper critical fields. The curves have been shifted with respect to one another for readability purposes. The arrows indicate the appearance of sharp kinks at higher temperatures. **c**, Specific heat of Br-DTN from $H = 0$ T to $H = 2$ T. **d**, Specific heat of Br-DTN in the Mott glass and Bose glass phases for $H \le H_{c1} \approx 1$ T, showing a non-exponential decay as $T \to 0$; a comparison is made to the predictions of theory based on the local-gap model (LGM), and to the data for pure DTN; in the upper-left and lower-right panels, the blue dashed line is a fit of the pure-DTN data to $A\exp(-\Delta(H)/k_BT)$, where A is a constant and $\Delta(H)\big/k_B = g\mu_B\left(H_{c1}^{(0)} - H\right)\big/k_B = 3.16$ K for $H = 0$ and 1.64 K for $H = 1$ T. Error bars, 1 s.d.

outside the BEC region is completely different to that of the pure system. In the pure system, the ground state outside the magnetic BEC is a Mott insulator with a large spin gap Δ away from the critical fields. This leads to an exponential suppression of the specific heat C_V at low temperatures $k_BT \lesssim \Delta$ as $C_V \propto \exp[-\Delta/(k_BT)]$, as shown in Fig. 2d, and to a similarly vanishing susceptibility for $T \to 0$. On the contrary, for $x = 0.08$, we observe that the susceptibility is finite for $H \ge H_{c2}$, and it even exhibits a strong satellite peak for $H \approx 13.5$ T. The susceptibility vanishes only for $H = H_s \approx 17$ T, corresponding to the saturation field of the entire sample, which is pushed to a much higher value than in the pure sample (where $H_s = H_{c2}^{(0)} = 12.6$ T). In the region $H \le H_{c1}$ we observe that the specific heat exhibits a non-exponential decay, down to zero field (Fig. 2d). Therefore we can conclude that the non-magnetic phases for $0 \le H \le H_s$ correspond to gapless bosonic insulators, which, as we will see, can be identified with a compressible Bose glass (for $H > 0$) and an incompressible Mott glass (for $H = 0$).

Modelling Br doping

Br-DTN can be successfully modelled theoretically by considering that Br substitution for Cl affects the super-exchange paths associated with the J_c couplings, and it also affects the crystal field locally owing to the larger atomic radius of Br with respect to Cl. The disappearance of the spin gap down to $H = 0$ and the upward shift of the saturation field suggests that Br doping locally strengthens the magnetic coupling J_c and lowers the anisotropy D. For simplicity, we only consider that Ni–Cl–Cl–Ni bonds in DTN can be turned into Ni–Cl–Br–Ni or Ni–Br–Cl–Ni, and we neglect Ni–Br–Br–Ni bonds that represent only 0.6% of the total bonds for $x = 0.08$.

We assign a J_c' value to the magnetic exchange coupling of the Br-doped bonds, and a D' value to the single-ion anisotropies of the Ni ion adjacent to the Br dopant. Note that for a doping concentration x, we have a fraction $2x$ of doped bonds, given that each bond can accommodate a Br dopant on two different Cl sites. We then use J_c' and D' as fitting parameters of the full low-temperature magnetization curve in Fig. 2a, which is calculated using QMC simulations (see Supplementary Information). We find an extremely good agreement between experimental data and simulation for $J_c' \approx 2.35 J_c$ and $D' \approx D/2$, giving us confidence that we are able to quantitatively model the fundamental microscopic effects of doping in Br-DTN. Indeed, the critical temperature for Bose–Einstein condensation, extracted from a finite-size scaling analysis of the simulation data with doping $x = 0.075$ (see Supplementary Information), is in

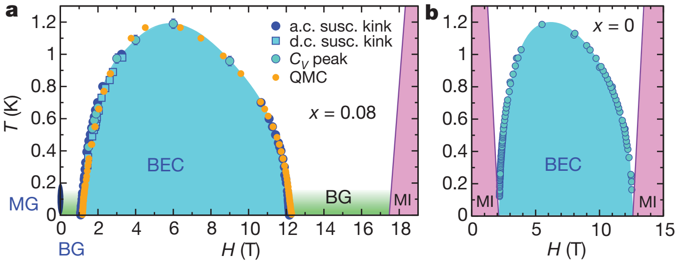


Figure 3 | Phase diagrams in the field–temperature plane. a, Experimental phase diagram of Br-doped DTN from specific heat and susceptometry, compared to QMC data. The following phases are represented: Bose-Einstein condensate (BEC), Bose glass (BG) and Mott glass (MG). The lilac regions represent the magnitude of the spin gap in the Mott insulating (MI) phase. **b**, Experimental phase diagram of pure DTN (based on specific heat and the magnetocaloric effect[22]).

remarkable quantitative agreement with the experiment, as shown in Fig. 3a. The critical fields estimated from simulations are $H_{c1} = 1.172(5)$ T and $H_{c2} = 12.302(5)$ T, slightly larger (by ~0.1 T and 0.14 T, respectively) than the experimental values. However the large downward shift of H_{c1} (by about 1 T) with respect to the pure system is correctly captured. In the following we discuss the main features of the Bose glass and Mott glass phases expected for the model of Br-DTN, and corroborate such expectations quantitatively with the experimental data.

Bose and Mott glasses

The results of our experiments on Br-DTN are consistent with a Bose glass for certain applied magnetic fields, and also with a Mott glass for $H = 0$ (see Fig. 3). A Mott glass has the peculiar property of being incompressible—that is, it has a vanishing susceptibility at $T = 0$, despite being gapless[19–21] (see also Supplementary Information). Br-DTN represents to our knowledge the first experimental realization of a Mott glass. To understand how the Mott glass applies to Br-DTN, consider that a compound in which all *c*-axis bonds contain a Br dopant (leading to couplings J_c' everywhere, and to an anisotropy D' on one of the two sites connected by the bond) will be a BEC even in zero field (see Supplementary Information). This means that rare Br-rich regions in Br-DTN behave locally as mini-BECs, and hence they are locally gapless. Strictly speaking, Br-rich regions will have a residual gap owing to their finite size. However, the statistical distribution of sizes has no upper bound, so that the distribution of local gaps has no lower bound, and consequently Br-DTN is globally gapless even in zero field. The corresponding bosonic phase is therefore a gapless insulator with spin inversion symmetry along the field axis, a commensurate boson density $n = 1$, and a vanishing compressibility resulting from the above symmetry[25]. As soon as a magnetic field is applied to this Mott glass, excess bosons are injected, which Anderson-localize around the rare Br-rich regions, resulting in a Bose glass (Fig. 1c). In the spin language, spins in the Br-rich regions acquire a finite magnetization along the field, and their transverse components correlate antiferromagnetically over a finite range, but the local phase of the antiferromagnetic order is different from region to region so that the system remains globally paramagnetic. Long-range phase coherence of the local order parameters—corresponding to the local phases of the bosonic wavefunction—is established only when the localized states of the bosons grow enough under the action of the applied field to overlap, leading to coherent tunnelling of bosons between neighbouring localized states (Fig. 1d). The resulting phase is a highly inhomogeneous BEC[26].

We can quantitatively test the picture of bosons localized in rare Br-rich regions against the thermodynamic behaviour of Br-DTN by using a simplified local-gap model (LGM). Within this model (see Supplementary Information), the low-temperature and low-field behaviour of the system is reduced to that of a collection of three-level systems, corresponding to a local longitudinal magnetization $m_{S,\mathrm{tot}} = 0, \pm 1$ for each localized state. There is a finite-size gap $\Delta_N \approx c/N$ (for zero field) between the $m_{S,\mathrm{tot}} = 0$ ground state and the $m_{S,\mathrm{tot}} = \pm 1$ excited states, where N is the number of sites in the Br-rich cluster and c is a fitting parameter. The low-temperature specific heat in zero field can then be predicted analytically to be

$$C_V(T) \propto t^{-5/4} \exp\left(-2\sqrt{c x_0 / t}\right) \qquad (2)$$

where $t = k_B T / J_c$ and $x_0 = \log(2x)$; this expression displays a stretched exponential behaviour that uniquely characterizes the Mott glass[21]. The c parameter, and an overall prefactor, are used as fitting parameters of the experimental data in zero field, leading to an extremely good fit, as shown in Figs 2d and 4. Notably, no further adjustable parameters are necessary to fit the finite-field data, displayed in Fig. 2d, which also show a remarkable agreement with the theory prediction up to $H \approx H_{c1}$.

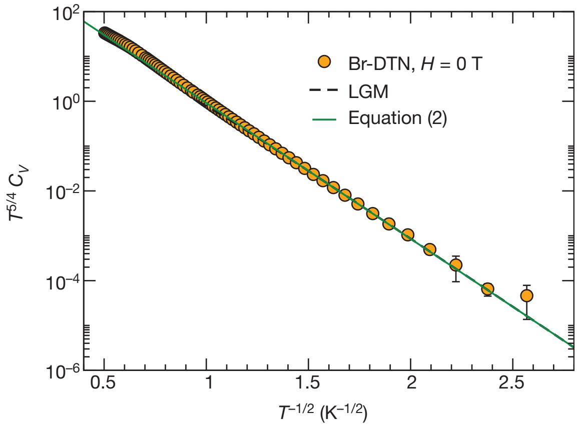


Figure 4 | Specific heat scaling and Mott glass. The specific heat in zero field is seen to display the characteristic Mott glass scaling at low temperatures, $\exp(-T^{-1/2})$. The solid lines are theoretical predictions based on the numerical solution of the local gap model (LGM) and its approximate analytical solution given by equation (2), with parameter $c = 3.02$. Error bars, 1 s.d.

High-field Bose glass

For $H \to H_{c2}$ the magnetization approaches the value $m_x = 1 - 2x \approx 0.84$, where all spins not connected to a Br-doped bond are polarized—and indeed H_{c2} lies very close to the polarization field $H_{c2}^{(0)}$ of pure DTN. The full polarization of the Br-poor regions leads to a pseudo-plateau in the magnetization at $m \approx m_x$ (pseudo because it still exhibits a small finite slope—a similar feature has also been observed in a Br-doped spin ladder at high field[14]). We interpret this feature as corresponding to the high-field Bose glass, which is characterized by the localization of bosonic holes, or singly occupied sites with $m_S = 0$, in a background of doubly occupied sites with $m_S = 1$ (Fig. 1f). The magnetically disordered nature of the high-field Bose glass phase could only be inferred from the susceptibility data and from our numerics. Indeed, in the experiments we could not determine unambiguously the absence of a λ-anomaly in the specific heat at low temperatures for $H \approx H_{c2}$, given that at such high fields the low-temperature specific heat of DTN is dominated by a Schottky anomaly whose origin can be ascribed to nuclear spins[27]. The localized bosonic holes persist up to the saturation field H_s, which is roughly the field necessary to fully polarize a homogeneous system with J_c' couplings and D' anisotropies everywhere. The step-like feature in the magnetization at the upper bound of the pseudo-plateau is therefore induced by the saturation of the Br-rich clusters, and it is smeared owing to the fact that such clusters have random geometries and therefore a distribution of local saturation fields, with an upper bound of H_s. One might suspect that the peak anomaly in the susceptibility corresponding to the step feature in the magnetization is associated with a further transition, but the numerical data, showing the same anomaly, allow us to conclude that the ground state is disordered in that field range.

Thermal percolation crossover

The physics described so far is valid only for very low temperatures. As the temperature is increased (above ~200 mK, as we will see below), the bosons that were localized in the Bose glass state are expected to thermally delocalize and proliferate. This leads to a thermal percolation of their density profile (corresponding to the longitudinal magnetization profile) throughout the sample[26]. Thus a more ordinary paramagnetic state is expected to appear at higher temperatures, and the nature of the field-driven transition into the BEC phase should also change fundamentally. Indeed, at temperatures below the thermal percolation crossover, the BEC transition should occur as sketched in Fig. 1c–f, by coherent tunnelling of bosons between localized states, resulting in a highly inhomogeneous BEC phase. This picture changes above the thermal percolation crossover. Now in the normal phase, the bosons move incoherently on a pre-percolated network of magnetized sites,

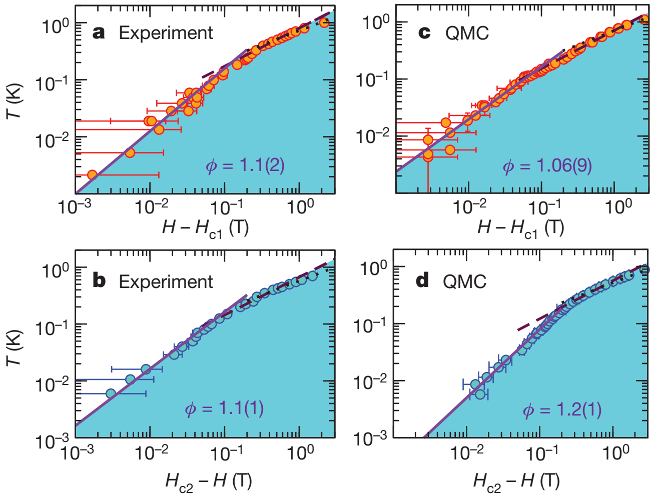


Figure 5 | Critical temperature scaling close to the zero-temperature critical fields. The scaling of the critical temperature with the distance from the $T = 0$ critical fields exhibits a crossover between various exponents. The dashed and dotted lines indicate a fit to the form $a|H - H_{c1(2)}|^{2/3}$ and $a|H - H_{c1(2)}|^{1/2}$ respectively, while the solid line is a fit to $a'|H - H_{c1(2)}|^{\phi}$, with the resulting ϕ exponent indicated in the figure (a and a' are fitting parameters). **a**, **b**, The critical line extracted from the a.c. susceptibility; **c**, **d**, the critical line obtained from the QMC simulations. Error bars, 1 s.d.

and their BEC transition upon increasing the field corresponds therefore to condensation on a random three-dimensional percolated lattice, which is fully analogous to condensation on a regular three-dimensional lattice. Signatures of the thermal percolation crossover can be found in the critical behaviour of the a.c. susceptibility: at low temperatures ($T \lesssim 200$ mK) it exhibits a rounded shoulder for $H \gtrsim H_{c1}$ and $H \lesssim H_{c2}$, and at higher temperatures it shows a sharp kink—analogous to what is observed in the pure system[17,24] (see Fig. 2b).

But the most marked signature of the thermal percolation crossover is observed in the scaling of the critical temperature with the applied field, shown in Fig. 5. Plotting T_c versus $|H - H_{c1(2)}|$ on a log–log scale (Fig. 5a, b), we clearly observe a kink separating two different scaling regimes. At high temperatures, ($T \gtrsim 200-300$ mK) the field-dependence of T_c is essentially consistent with a pure-system scaling for low temperatures, $T_c \propto |H - H_{c1(2)}|^{\phi}$ with $\phi = 2/3$, or with a pure-system scaling for intermediate temperatures with $\phi \approx 1/2$, as observed in other magnetic BEC systems[28]. At low temperatures, the scaling exponent crosses over to novel values, $\phi = 1.1(2)$ (close to H_{c1}) and $\phi = 1.1(1)$ (close to H_{c2}), which are consistent within the error (see Supplementary Information for a discussion of the estimate of ϕ). Moreover, these scaling exponents are also consistent with the values extracted from our QMC simulations (Fig. 5c, d; $\phi = 1.06(9)$ and $1.2(1)$ close to H_{c1} and H_{c2} respectively). Simulations also show a rough quantitative agreement for the crossover temperature range. Most remarkably, a consistent value of the exponent ϕ at low temperature is also observed theoretically for the magnetic Hamiltonian of DTN subject to a different type of disorder, namely site dilution[26]. We can therefore conclude that the low-temperature scaling of T_c exhibits an exponent $\phi \approx 1$–1.1 which is a universal feature of the Bose glass–BEC transition. Its value deviates from the prediction $\phi > 2$ of ref. 4, but this prediction is based on a scaling Ansatz for the free energy close to the quantum critical point which is found to be inconsistent with other observations on Br-DTN, as well as with numerical simulations[29]. Therefore we conclude that our results call for a generalization of the scaling assumptions for the disordered-boson quantum critical point.

Conclusions

We have performed a comprehensive experimental and theoretical study of the disordered and strongly interacting Bose fluid realized in a doped quantum magnet (Br-DTN) under application of a magnetic field. We provide substantial evidence for the existence of gapless insulating phases of the bosons—the Mott glass and the Bose glass—and we investigate the quantitative features associated with their thermodynamic behaviour. These phases can be quantitatively described as a Bose fluid fragmented over an extensive number of localized states with variable local gaps, dominating the response of the system. The presence of a Bose glass leads to a novel and seemingly universal exponent governing the scaling of the critical temperature for the transition from Bose glass to BEC. The remarkable agreement between theory and experiment shows that Br-DTN is an extremely well controlled realization of a disordered Bose fluid, which allows a detailed experimental study of the thermal phase diagram of disordered bosons in the grand-canonical ensemble.

METHODS SUMMARY

Br-DTN crystals were prepared at the University of São Paulo, and their X-ray analysis was performed at the Los Alamos National Laboratories. The same crystal was used for a.c. susceptibility and specific heat measurements—the specific heat sample was a small slice of the a.c. susceptibility sample. All measurements were made with the magnetic field applied along the tetragonal axis (c axis) of the sample. The a.c. susceptibility measurements were carried out using a $PrNi_5$ nuclear refrigerator (down to 1 mK) and a 15 T magnet at the High B/T facility of the National High Magnetic Field Laboratory in Gainesville. The field sweep rates were adjusted to values as low as 10^{-3} T min^{-1} to guarantee the full relaxation of the sample at the lowest temperatures probed. The d.c. magnetization was measured at the University of São Paulo by using a dilution refrigerator at 19 mK. Specific heat was measured in a Quantum Design ^{3}He/^{4}He dilution refrigerator down to 50 mK using the thermal relaxation method. The numerical simulations were based on the stochastic series expansion approach with directed-loop updates, and they were performed on the Jaguar cluster of the National Center for Computational Sciences (Oak Ridge National Laboratories).

Received 14 May; accepted 11 July 2012.

1. Kramer, B. & MacKinnon, A. Localization: theory and experiment. *Rep. Prog. Phys.* **56,** 1469–1564 (1993).
2. Fallani, L., Fort, C. & Inguscio, M. Bose-Einstein condensates in disordered potentials. *Adv. At. Mol. Opt. Phys.* **56,** 119–160 (2008).
3. Giamarchi, T. & Schulz, H. J. Anderson localization and interactions in one-dimensional metals. *Phys. Rev. B* **37,** 325–340 (1988).
4. Fisher, M. P. A., Weichman, P. B., Grinstein, G. & Fisher, D. S. Boson localization and the superfluid-insulator transition. *Phys. Rev. B* **40,** 546–570 (1989).
5. Crowell, P. A., Van Keulz, F. W. & Reppy, J. D. Onset of superfluidity in ^{4}He films adsorbed on disordered substrates. *Phys. Rev. B* **55,** 12620–12634 (1997).
6. Sacépé, B. *et al.* Localization of preformed Cooper pairs in disordered superconductor. *Nature Phys.* **7,** 239–244 (2011).
7. Bouadim, K., Loh, Y. L., Randeria, M. & Trivedi, N. Single- and two-particle energy gaps across the disorder-driven superconductor-insulator transition. *Nature Phys.* **7,** 884–889 (2011).
8. Sanchez-Palencia, L. & Lewenstein, M. Disordered quantum gases under control. *Nature Phys.* **6,** 87–95 (2010).
9. Deng, H. Haug, H. & Yamamoto, Y. Exciton-polariton Bose-Einstein condensation. *Rev. Mod. Phys.* **82,** 1489–1537 (2010).
10. Giamarchi, T., Rüegg, Ch. & Tchernyshyov, O. Bose-Einstein condensation in magnetic insulators. *Nature Phys.* **4,** 198–204 (2008).
11. Nohadani, O., Wessel, S. & Haas, S. Bose-glass phases in disordered quantum magnets. *Phys. Rev. Lett.* **95,** 227201 (2005).
12. Roscilde, T. & Haas, S. Quantum localization in bilayer Heisenberg antiferromagnets with site dilution. *Phys. Rev. Lett.* **95,** 207206 (2005).
13. Roscilde, T. Field-induced quantum-disordered phases in S = 1/2 weakly coupled dimer systems with site dilution. *Phys. Rev. B* **74,** 144418 (2006).
14. Manaka, H., Kolomiets, A. V. & Goto, T. Disordered states in IPA-$Cu(Cl_xBr_{1-x})_3$ induced by bond randomness. *Phys. Rev. Lett.* **101,** 077204 (2008).
15. Hong, T., Zheludev, A., Manaka, H. & Regnault, L.-P. Evidence of a magnetic Bose glass in $(CH_3)_2CHNH_3Cu(Cl_{0.95}Br_{0.05})_3$ from neutron diffraction. *Phys. Rev. B* **81,** 060410 (2010).
16. Zapf, V. S. *et al.* Bose-Einstein condensation of $S = 1$ nickel spin degrees of freedom in $NiCl_2$-4SC$(NH_2)_2$. *Phys. Rev. Lett.* **96,** 077204 (2006).
17. Yin, L., Xia, J. S., Zapf, V. S., Sullivan, N. S. & Paduan-Filho, A. Direct measurement of the Bose-Einstein condensation universality class in $NiCl_2$-4SC$(NH_2)_2$ at ultralow temperatures. *Phys. Rev. Lett.* **101,** 187205 (2008).
18. Orignac, E., Giamarchi, T. & Le Doussal, P. A possible new phase of commensurate insulators with disorder: the Mott glass. *Phys. Rev. Lett.* **83,** 2378–2381 (1999).
19. Prokof'ev, N. & Svistunov, B. Superfluid-insulator transition in commensurate disordered bosonic systems: large-scale worm algorithm simulations. *Phys. Rev. Lett.* **92,** 015703 (2004).
20. Altman, E., Kafri, Y., Polkovnikov, A. & Refael, G. Phase transition in a system of one-dimensional bosons with strong disorder. *Phys. Rev. Lett.* **93,** 150402 (2004).

21. Roscilde, T. & Haas, S. Mott glass in site-diluted S=1 antiferromagnets with single-ion anisotropy. *Phys. Rev. Lett.* **99,** 047205 (2007).
22. Zvyagin, S. A. *et al.* Magnetic excitations in the spin-1 anisotropic Heisenberg antiferromagnetic chain system $NiCl_2$-4SC$(NH_2)_2$. *Phys. Rev. Lett.* **98,** 047205 (2007).
23. Zvyagin, S. A. *et al.* Spin dynamics of $NiCl_2$-4SC$(NH_2)_2$ in the field-induced ordered phase. *Phys. Rev. B* **77,** 092413 (2008).
24. Yin, L., Xia, J. S., Zapf, V. S., Sullivan, N. S., &. Paduan-Filho, A. Magnetic susceptibility measurements at ultra-low temperatures. *J. Low Temp. Phys.* **158,** 710–715 (2010).
25. Balabanyan, K. G., Prokof'ev, N. & Svistunov, B. Superfluid-insulator transition in commensurate one-dimensional bosonic system with off-diagonal disorder. *Phys. Rev. Lett.* **95,** 055701 (2005).
26. Yu, R., Haas, S. & Roscilde, T. Universal phase diagram of disordered bosons from a doped quantum magnet. *Europhys. Lett.* **89,** 10009 (2010).
27. Weickert, F. *et al.* Low temperature thermodynamic properties near the field-induced quantum critical point in DTN. *Phys. Rev. B* **85,** 184408 (2012).
28. Kawashima, N. Quantum critical point of the *XY* model and condensation of field-induced quasiparticles in dimer compounds. *J. Phys. Soc. Jpn* **73,** 3219–3222 (2004).
29. Yu, R. *et al.* Quantum critical scaling at a Bose-glass/superfluid transition: theory and experiment on a model quantum magnet. Preprint at http://arXiv.org/abs/1204.5409 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements Work at the High Magnetic Field Laboratory at the Physics Institute of the University of São Paulo was supported in part by the Brazilian agencies FAPESP and CNPq. Measurements at the NHMFL High B/T and pulsed field facilities were supported by NSF grant DMR 0654118, by the State of Florida, and by the DOE. Work at LANL was supported by the NSF, and by the DOE's Laboratory Directed Research and Development programme under 20100043DR. C.F.M. acknowledges support by UEFISCDI (project RP-10). The numerical simulations were performed on the computer facilities of the NCCS at the Oak Ridge National Laboratories, and supported by INCITE Award MAT013 of the Office of Science, DOE. Further theory work was supported by the DOE (grant DE-FG03-01ER45908 and DE-FG02-05ER46240), the NSF (grant DMR-1006985) and by the Robert A. Welch Foundation (grant C-1411).

Author Contributions R.Y., S.H. and T.R. performed the numerical simulations and the theoretical analysis; L.Y., N.S.S., J.S.X. and C.H. performed the susceptibility measurements; A.P.-F. and N.F.O. synthesized the samples and performed the d.c. magnetization measurements; A.S., C.F.M., F.W., R.M., E.-D.M. and V.S.Z. performed the specific heat measurements; and B.L.S. performed the X-ray measurements. V.S.Z. coordinated the experimental efforts and T.R. the theoretical ones. T.R. wrote the manuscript, with the contributions of all the authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R. (tommaso.roscilde@ens-lyon.fr).

doi:10.1038/nature11356

Autistic-like behaviour in *Scn1a*$^{+/-}$ mice and rescue by enhanced GABA-mediated neurotransmission

Sung Han[1,2,3], Chao Tai[2], Ruth E. Westenbroek[2], Frank H. Yu[2]†, Christine S. Cheah[2], Gregory B. Potter[4], John L. Rubenstein[4], Todd Scheuer[2], Horacio O. de la Iglesia[1,3] & William A. Catterall[1,2]

Haploinsufficiency of the *SCN1A* gene encoding voltage-gated sodium channel Na$_V$1.1 causes Dravet's syndrome, a childhood neuropsychiatric disorder including recurrent intractable seizures, cognitive deficit and autism-spectrum behaviours. The neural mechanisms responsible for cognitive deficit and autism-spectrum behaviours in Dravet's syndrome are poorly understood. Here we report that mice with *Scn1a* haploinsufficiency exhibit hyperactivity, stereotyped behaviours, social interaction deficits and impaired context-dependent spatial memory. Olfactory sensitivity is retained, but novel food odours and social odours are aversive to *Scn1a*$^{+/-}$ mice. GABAergic neurotransmission is specifically impaired by this mutation, and selective deletion of Na$_V$1.1 channels in forebrain interneurons is sufficient to cause these behavioural and cognitive impairments. Remarkably, treatment with low-dose clonazepam, a positive allosteric modulator of GABA$_A$ receptors, completely rescued the abnormal social behaviours and deficits in fear memory in the mouse model of Dravet's syndrome, demonstrating that they are caused by impaired GABAergic neurotransmission and not by neuronal damage from recurrent seizures. These results demonstrate a critical role for Na$_V$1.1 channels in neuropsychiatric functions and provide a potential therapeutic strategy for cognitive deficit and autism-spectrum behaviours in Dravet's syndrome.

Dravet's syndrome (DS), also called severe myoclonic epilepsy of infancy, is an intractable developmental epilepsy syndrome with seizure onset in the first year of life[1]. However, unlike other generalized epilepsy disorders, it is accompanied by characteristic neuropsychiatric comorbidities, including hyperactivity, attention deficit, delayed psychomotor development, sleep disorder, anxiety-like behaviours, impaired social interactions, restricted interests and severe cognitive deficits[1–6]. These comorbidities in DS overlap with symptoms of autism-spectrum disorders (ASD), and a recent study suggests that DS patients have autism-spectrum behaviours[3]. DS is caused by heterozygous loss-of-function mutations in the *SCN1A* gene[7], which encodes the pore-forming α-subunit of the brain voltage-gated sodium channel type-1 (Na$_V$1.1)[8]. As in DS, mice with heterozygous loss-of-function mutation in *Scn1a* (*Scn1a*$^{+/-}$) have thermally induced and spontaneous seizures, premature death, ataxia and sleep disorder[9–13]. Na$_V$1.1 channels are expressed in cell bodies and axon initial segments of excitatory and inhibitory neurons in the brain[14–16], but deletion of Na$_V$1.1 impairs Na$^+$ currents and action potential firing of GABAergic interneurons specifically because Na$_V$1.1 is the primary Na$^+$ channel in those cells[9,16]. Specific deletion of Na$_V$1.1 channels in forebrain interneurons using a Cre-LoxP strategy recapitulates the symptoms of DS in mice[17], confirming that loss of Na$_V$1.1 in GABAergic interneurons causes this disease. Emerging genetic evidence implicates *SCN1A* in autism[18–22], and there is increasing evidence that dysfunction of GABAergic signalling is associated with ASDs[23–25], leading to the proposal that elevation of excitation/inhibition ratio in neocortical neurons is the primary cause of ASD[26–29]. In this study, we have investigated autism-related behaviours in *Scn1a*$^{+/-}$ mice and shown that they are caused by impaired GABAergic neurotransmission that can be rescued by drug treatment.

Hyperactivity, anxiety and stereotypies in *Scn1a*$^{+/-}$ mice

Homozygous *Scn1a*$^{-/-}$ mice developed severe ataxia and died on postnatal day (P) 15, whereas *Scn1a*$^{+/-}$ mice had spontaneous seizures and sporadic deaths beginning after P21 (ref. 9). *Scn1a*$^{+/-}$ mice develop multiple behavioural phenotypes, which are phenocopies of comorbidities in DS. During a 10-min open-field test, adult *Scn1a*$^{+/-}$ mice travelled significantly farther than wild type (Fig. 1a), but spent less time in the centre of the open field (Fig. 1b and Supplementary Fig. 1). *Scn1a*$^{+/-}$ mice also spent more time self-grooming than wild type (Fig. 1c and Supplementary Fig. 3a) and showed increased circling behaviour (Fig. 1d and Supplementary Fig. 3b). In the elevated plus maze, *Scn1a*$^{+/-}$ mice entered open arms less frequently compared with wild type (Fig. 1e), and spent less time in the open arms (Fig. 1f and Supplementary Fig. 2). These observations indicate that *Scn1a*$^{+/-}$ mice exhibit hyperactivity, increased anxiety and increased stereotyped behaviours, which are phenocopies of autistic traits in DS. *Scn1a*$^{+/-}$ mice also have decreased nest-building ability compared to wild type (Supplementary Fig. 4), which could indicate deficits in social behaviour[30].

Scn1a$^{+/-}$ mice have deficits in social interaction

We performed behavioural tests to assess deficits in social interaction, a prominent symptom of ASD[31]. A three-chamber experiment showed that *Scn1a*$^{+/-}$ mice have profound deficits in social interaction. Both *Scn1a*$^{+/-}$ and wild type had no preference for two empty cages, located in the right and the left chambers during a habituation period (Supplementary Figs 5, 6, 7a). However, when we put a stranger mouse in the cage in one chamber, wild-type mice spent more time in the mouse-containing chamber than in the empty cage-containing chamber (Fig. 1g and Supplementary Fig. 5), and

[1] Graduate Program in Neurobiology & Behavior, University of Washington, Seattle, Washington 98195, USA. [2] Department of Pharmacology, University of Washington, Seattle, Washington 98195, USA. [3] Department of Biology, University of Washington, Seattle, Washington 98195, USA. [4] Department of Psychiatry, University of California at San Francisco, San Francisco, California 94158, USA. †Present address: Program in Neurobiology, School of Dentistry and Dental Research Institute, Seoul National University, Seoul 110-749, Korea.

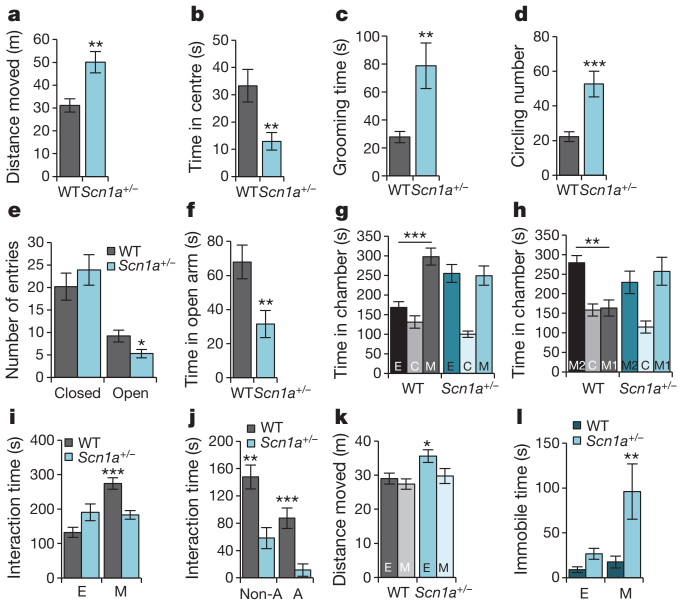


Figure 1 | *Scn1a*$^{+/-}$ mice show hyperactivity, anxiety-like behaviour, increased stereotypies, and impaired social behaviour. **a**, **b**, In the open field test, *Scn1a*$^{+/-}$ mice travel longer distances compared with wild-type mice (**a**) and spend less time in the centre (**b**). **c**, **d**, In the open field, *Scn1a*$^{+/-}$ mice spend more time grooming (**c**) and circling (**d**) than wild-type mice. In **d**, one complete turn is counted as one circle, regardless of direction. **e**, **f**, In the elevated plus maze, *Scn1a*$^{+/-}$ mice enter less frequently in the open arms (**e**) and spend less time in the open arms (**f**). **g**, **h**, Three-chamber experiment. **g**, Whereas wild-type mice spend more time in the chamber housing a stranger mouse (M) than the chamber housing an empty cage (E), *Scn1a*$^{+/-}$ mice have no preference for either chamber. **h**, Whereas wild-type mice spend more time in the chamber housing a novel mouse (M2) than in a chamber housing a familiar mouse (M1), *Scn1a*$^{+/-}$ mice have no preference for either chamber. **i**–**l**, Social interaction test. **i**, *Scn1a*$^{+/-}$ mice show decreased interaction with a caged stranger mouse when compared with wild-type mice. **j**, In a 10-min reciprocal interaction test, pairs of wild-type and *Scn1a*$^{+/-}$ unfamiliar mice had significantly less non-aggressive (Non-A) and aggressive (A) interactions than pairs of wild-type and wild-type unfamiliar mice. Aggressive behaviours included attacking, wrestling and biting the dorsal surface, and non-aggressive behaviours include nose-to-nose sniffing, anogenital sniffing and grooming. **k**, *Scn1a*$^{+/-}$ mice move significantly less when they encountered the stranger mouse compared to an empty cage, whereas there is no difference in movement for wild type. **l**, *Scn1a*$^{+/-}$ mice, but not wild-type mice, show increased immobilization behaviour in the presence of the caged stranger mouse than in the presence of an empty cage. All data shown are means ± s.e.m. from 10–12 mice per genotype. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

interacted more extensively with peer mice than with the empty cage (Supplementary Fig. 7b). In contrast, *Scn1a*$^{+/-}$ mice showed no preference for the stranger mouse (Fig. 1g and Supplementary Figs 5 and 7b). When a second stranger mouse was placed in the unoccupied side chamber to assess the discrimination between a new and a familiar mouse, wild-type mice showed strong preference for the new mouse, but *Scn1a*$^{+/-}$ mice did not (Fig. 1h and Supplementary Figs 5 and 7c), even though they have preference for new objects (see below). We observed similar social deficits of *Scn1a*$^{+/-}$ mice in the open field social interaction test. *Scn1a*$^{+/-}$ mice interacted significantly less with a caged stranger mouse in an open field compared with wild type mice (Fig. 1i and Supplementary Fig. 8a). When both the inanimate object and the caged stranger mouse were introduced simultaneously, wild-type mice interacted significantly more with a caged stranger mouse than with an empty cage, whereas *Scn1a*$^{+/-}$ mice showed no preference for the caged mouse (Supplementary Fig. 9a, b). We also examined reciprocal social interactions of freely moving *Scn1a*$^{+/-}$ and wild-type littermates with test mice. *Scn1a*$^{+/-}$ mice showed decreased duration of both non-aggressive and aggressive interactions (Fig. 1j). We observed that *Scn1a*$^{+/-}$ mice exhibited increased immobilization behaviour when they encountered the caged stranger mouse (Supplementary Fig. 8b). Compared to wild type, this immobilization decreased distance travelled (Fig. 1k) and increased immobilization time by 400% (Fig. 1l). Taken together, these results indicate that *Scn1a*$^{+/-}$ mice have profound deficits in social behaviour.

In nocturnal rodents, social interaction and olfactory perception are tightly associated[32], and impairment of olfactory perception leads to decreased social interaction[33]. We assessed olfaction in modified three-chamber experiments in which a tightly sealed Petri dish containing food pellets and an identical one with holes were placed in the side chambers. Both *Scn1a*$^{+/-}$ and wild-type mice spent more time in the food-odour chamber, showed a shorter latency to enter it, and entered it more frequently than the odourless chamber (Supplementary Fig. 10a–d). Alternatively, we used bedding from male or female cages as a social odour. Wild-type mice had a strong preference for the chamber containing bedding, whereas *Scn1a*$^{+/-}$ mice had no preference for these social odours (Supplementary Fig. 11a, b, d, e). In close-interaction analysis, *Scn1a*$^{+/-}$ mice avoided interacting with male social cues (Supplementary Fig. 11c), and both wild-type and *Scn1a*$^{+/-}$ mice strongly avoided fox urine (Supplementary Fig. 11f). Wild-type mice exhibited strong habituation and dishabituation to odours of banana, male urine and standard food, whereas *Scn1a*$^{+/-}$ mice gave a normal response to food but failed to show habituation/dishabituation to banana or male urine (Supplementary Fig. 12a). However, *Scn1a*$^{+/-}$ mice had greatly increased digging behaviour when banana and male urine odours were presented, indicating that they detect these odours (Supplementary Fig. 12b). Moreover, in a Y-maze olfactory choice test, *Scn1a*$^{+/-}$ mice strongly avoided banana and male urine, whereas wild-type mice had a strong preference for both (Supplementary Fig. 12c, d). These data indicate that *Scn1a*$^{+/-}$ mice perceive food odours and social olfactory cues, but they have no interest or avoid unfamiliar odours and social odours. These results further establish a deficit in social interaction[34,35] and avoidance of environmental change[36] in *Scn1a*$^{+/-}$ mice, as in ASDs.

Scn1a$^{+/-}$ mice have deficits in context-dependent spatial memory

Both wild-type and *Scn1a*$^{+/-}$ mice had similar ability to recognize a new object 24 h after training (Fig. 2a, b). In the context-dependent fear-conditioning test, *Scn1a*$^{+/-}$ and wild-type mice showed no freezing behaviour during the habituation period in context, and both of them had similar freezing behaviour immediately after a mild foot shock (Fig. 2c). However, whereas wild-type mice showed sustained freezing behaviours when returned to the shock cage 30 min and 24 h later, *Scn1a*$^{+/-}$ mice had substantially reduced freezing behaviour (Fig. 2c). The loss of fear-associated freezing behaviour was specific because measurements of distance and velocity of movement during the fear-conditioning test did not reveal other fear-associated responses such as panic fleeing (Supplementary Fig. 13).

To assess spatial learning and memory in the absence of fear, we performed the Barnes circular maze test in which mice learn to rapidly escape a brightly lighted circular field by finding a specific dark hole at its periphery. *Scn1a*$^{+/-}$ mice failed to improve their learning performance during four days of training (Fig. 2d, e), and had substantially reduced spatial memory during the probe trials at day 5 (Fig. 2f–i). These data, together with the results of the context-dependent fear-conditioning test (Fig. 2c), indicate that *Scn1a*$^{+/-}$ mice have severely impaired spatial learning and memory.

Conditional *Scn1a*$^{+/-}$ mutant mice exhibit autism-related behaviours

To determine whether the autism-related phenotypes of *Scn1a*$^{+/-}$ mice emerge specifically from reduced Na$_V$1.1 activity in forebrain GABAergic neurons, we generated forebrain GABAergic neuron-specific conditional *Scn1a*$^{+/-}$ mutant mice using the *DlxI12b-Cre*

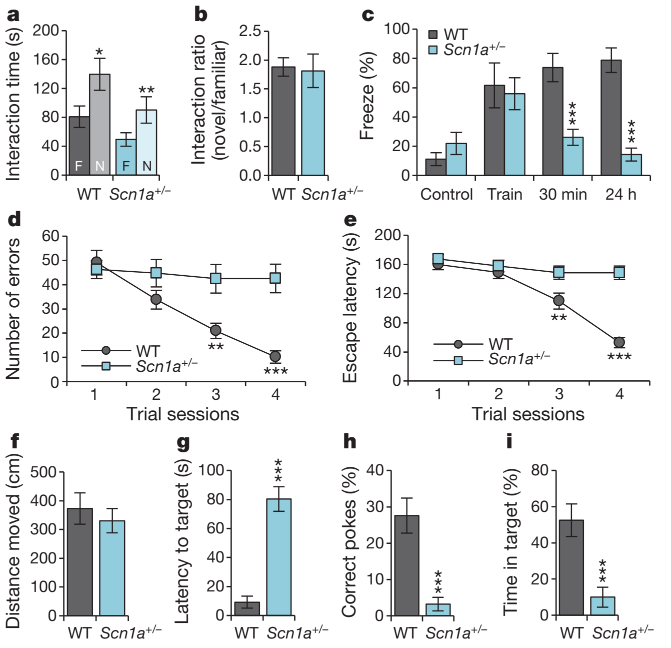


Figure 2 | Profound deficits in context-dependent spatial learning and memory in $Scn1a^{+/-}$ mice. **a**, In the novel object recognition test, $Scn1a^{+/-}$ mice had normal recognition memory for a preconditioned object (F: Familiar), which was presented 24 h before the test, so that they spent more time with the novel object (N: Novel). **b**, Discrimination index, the normalized ratio of time spent with the familiar object divided by time spent with the novel object, shows that there is no difference between wild-type and $Scn1a^{+/-}$ mice for novel object recognition ability. **c**, In the contextual fear-conditioning test, $Scn1a^{+/-}$ mice had a normal fear response immediately after the training (Train), but showed a profound deficit in short-term (30 min) and long-term (24 h) memory for the spatial context associated with a 2-s mild foot shock (0.5 mA) when compared to wild-type mice. **d**, **e**, In the Barnes circular maze, $Scn1a^{+/-}$ mice had a profound deficit in spatial learning. Wild-type mice made fewer errors finding the target hole (**d**), and showed decreased latency to escape the maze (**e**) during the 4-day repeated training trials, but $Scn1a^{+/-}$ mice show no improvement in performance for either the number of errors made to find the target hole (**d**), or the time to escape the maze (**e**). **f–i**, During the probe trial on the 5th day, $Scn1a^{+/-}$ mice had a profound deficit in spatial memory. They spent significantly more time to find the target hole (**g**), poked the correct target hole with significantly lower frequency (**h**), and stayed significantly less time in the target area (**i**) when compared with wild-type mice, although total distance moved was not significantly different from that of wild-type mice (**f**). All data shown are means ± s.e.m. from 6–10 mice per genotype. $^{*}P < 0.05$, $^{***}P < 0.001$.

Cre-recombinase mouse line (*Dlx1/2-Cre*[17,37,38]). These mice have a specific reduction of $Na_V1.1$ channels in forebrain GABAergic neurons and have similar epilepsy and premature death as $Scn1a^{+/-}$ mice[17]. $Dlx1/2^{+}$ *Scn1a* heterozygous mutant mice ($Dlx1/2\text{-}Scn1a^{+/-}$) recapitulated the autism-related phenotypes and spatial learning deficit of $Scn1a^{+/-}$ mice (Fig. 3), whereas control Cre-positive $Scn1a^{+/+}$ mice did not (Supplementary Fig. 14). In the open field test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice travelled more (Fig. 3a), spent less time in the centre (Fig. 3b), and showed increased circling behaviour (Fig. 3c)[17] when compared with Cre-negative $Scn1a^{+/loxp}$ mice. In the elevated-plus maze test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice entered the open arms less frequently and spent less time in open arms than Cre-negative $Scn1a^{+/loxp}$ mice (Fig. 3d, e). In the open field social interaction test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice spent less time interacting with the caged stranger mouse compared to Cre-negative $Scn1a^{+/loxp}$ mice (Fig. 3f). In addition, in the three-chamber social preference test Cre-negative $Scn1a^{+/loxp}$ mice stayed longer in the mouse chamber versus the inanimate-object chamber; in contrast, $Dlx1/2\text{-}Scn1a^{+/-}$ mice showed no preference (Fig. 3g). Finally, in the contextual fear-conditioning test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice showed similar freezing behaviour in control and training sessions, but significantly less freezing behaviour in the 30 min and 24 h after the training compared with $Scn1a^{+/loxp}$ mice (Fig. 3h). These results show that $Dlx1/2\text{-}Scn1a^{+/-}$ mice reproduce hyperactive and anxiety-like behaviours, deficits in social interactions, and impaired context-dependent fear conditioning of global $Scn1a^{+/-}$ mice. This evidence indicates that the autism-related phenotype emerges from reduced $Na_V1.1$ activity specifically within forebrain GABAergic interneurons.

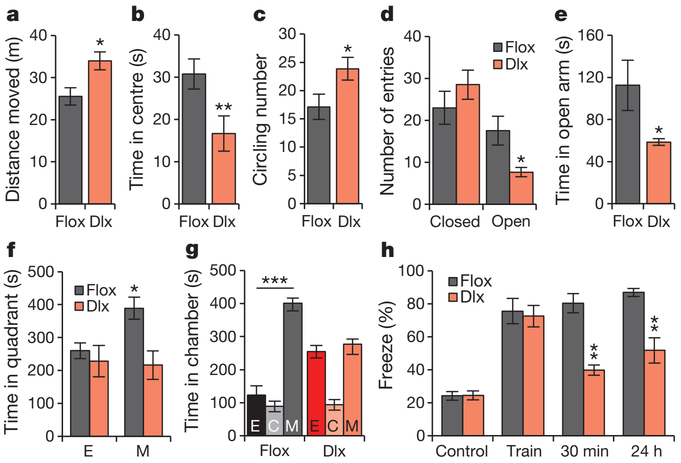


Figure 3 | $Dlx1/2\text{-}Scn1a^{+/-}$ mice have the impaired spatial learning and autism-related phenotypes observed in $Scn1a^{+/-}$ mice. **a**, In the open field test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice moved farther compared to Cre-negative $Scn1a^{+/loxp}$ littermates. **b**, In the open field test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice spent less time in the centre. **c**, $Dlx1/2\text{-}Scn1a^{+/-}$ mice show increased circling behaviour. One complete turn, regardless of direction was counted as one circling. **d**, **e**, In the elevated plus maze, $Dlx1/2\text{-}Scn1a^{+/-}$ mice entered less frequently into open arms (**d**), and spent significantly less time in open arms (**e**). **f**, In the open field social interaction test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice showed decreased interaction with social cues compared to $Scn1a^{+/loxp}$ littermates. **g**, In the 3-chamber test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice had no preference for the stranger mouse. **h**, In the contextual fear conditioning test, $Dlx1/2\text{-}Scn1a^{+/-}$ mice had a normal fear response immediately after the foot shock during training but showed a profound deficit in short-term (30 min) and long-term (24 h) memory for the spatial context associated with a 2-s mild foot shock (0.5 mA) when compared to wild-type mice. Dlx, $Dlx1/2\text{-}Scn1a^{+/-}$ mice. Flox, Cre-negative $Scn1a^{+/loxp}$ mice. E, Empty cage. C, Center. M, Mouse. All data shown are means ± s.e.m. from 7–9 mice per genotype. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Deficit of $Na_V1.1$ channels impairs GABAergic neurotransmission

To test our hypothesis that the autism-related phenotypes and spatial learning deficits in $Scn1a^{+/-}$ mice are caused by decreased $Na_V1.1$ activity in GABAergic interneurons in the forebrain, we compared the properties of cortical and hippocampal GABAergic interneurons in wild-type and $Scn1a^{+/-}$ mice. $Na_V1.1$ protein is expressed in adult hippocampal and neocortical interneurons, as assessed by co-immunolabelling of $Na_V1.1$ channels and GABA in the hippocampal CA1 region (Fig. 4a) and prefrontal cortex (Supplementary Fig. 15). The proportion of GABAergic interneurons expressing a detectable level of $Na_V1.1$ in of $Scn1a^{+/-}$ mice was decreased 20–50% throughout the cortex and hippocampus (Fig. 4b), whereas there was no reduction in the total number of GABA-stained interneurons (Supplementary Fig. 16, legend). The deep layer of prefrontal cortex was the most affected by the *Scn1a* mutation (Fig. 4b), and the intensity of immunostaining for $Na_V1.1$ in GABAergic cells with detectable staining was reduced by 50% in the prefrontal cortex (Supplementary Fig. 16).

Some forms of autism are postulated to be caused by an imbalance of synaptic transmission between excitatory and inhibitory circuits[26–29]. $Scn1a^{+/-}$ mice have reduced Na^{+} currents and impaired action potential firing in both hippocampal interneurons and cerebellar Purkinje neurons[9,10], which are GABAergic neurons. When action

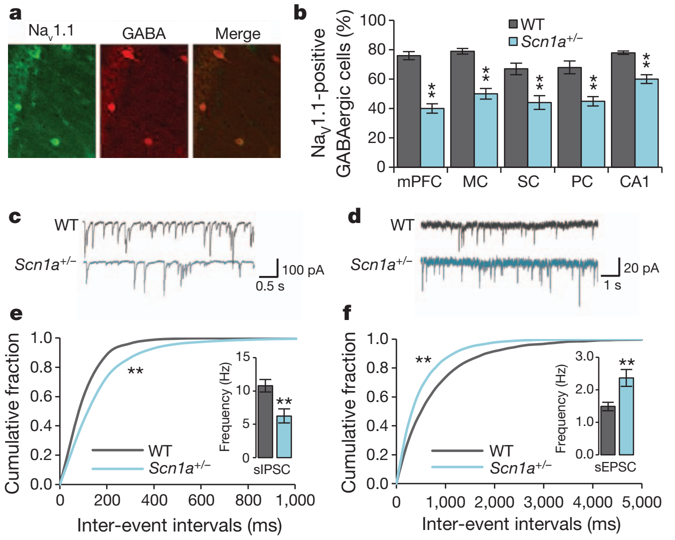


Figure 4 | Deficit of Na$_V$1.1 channels and GABAergic neurotransmission in *Scn1a*$^{+/-}$ hippocampal GABAergic interneurons. Immunocytochemical staining of forebrain neurons from 10-month old mice for Na$_V$1.1 channels. **a**, Co-immunolabelling of Na$_V$1.1 and GABA revealed co-expression of Na$_V$1.1 and GABA in the hippocampal CA1 region in wild-type mice. **b**, Co-immunolabelling of Na$_V$1.1 and GABA revealed decreased expression of Na$_V$1.1 channels in GABAergic interneurons in the forebrain of *Scn1a*$^{+/-}$ mice. **c**, Example traces of sIPSC from wild-type and *Scn1a*$^{+/-}$ hippocampal CA1 neurons. **d**, Example traces of sEPSC from wild-type and *Scn1a*$^{+/-}$ hippocampal CA1 neurons. **e**, Cumulative plot and average values (inset) of sIPSC frequency. The frequency of sIPSC is decreased, but the amplitude of sIPSC is unchanged in *Scn1a*$^{+/-}$ hippocampal CA1 slices when compared to wild-type slices (Supplementary Fig. 17a). **f**, Cumulative plot and average values (inset) of sEPSC frequency. The frequency of sEPSC is increased, but the amplitude of sEPSC is unchanged in *Scn1a*$^{+/-}$ hippocampal CA1 slices when compared to wild-type slices (Supplementary Fig. 17b). mPFC, medial prefrontal cortex. MC, motor cortex. SC, sensory cortex. PC, parietal cortex. CA1, hippocampal CA1 region. All data shown are means ± s.e.m. from 15–19 recordings per genotype. $^{**}P < 0.01$.

potentials were blocked with tetrodotoxin (TTX, 1 μM), recordings of miniature inhibitory postsynaptic currents (IPSC) and miniature excitatory postsynaptic current (EPSC) from the hippocampal CA1 region and the prefrontal cortex showed that amplitude and frequency were not altered, indicating normal synaptic function in *Scn1a*$^{+/-}$ slices (Supplementary Figs 17 and 18). Similarly, in the absence of TTX, the amplitudes of spontaneous IPSCs and spontaneous EPSCs were unchanged (Fig. 4c, d and Supplementary Figs 19 and 20), indicating that the postsynaptic response to released neurotransmitter was not altered. In contrast, in the absence of TTX, the frequency of spontaneous IPSCs in hippocampal CA1 and prefrontal cortex slices from *Scn1a*$^{+/-}$ mice was reduced (Fig. 4c, e and Supplementary Fig. 20a, b), and the frequency of spontaneous EPSCs was increased (Fig. 4d, f and Supplementary Fig. 20c, d) compared to wild-type slices. Because no differences in frequencies of miniature IPSCs or EPSCs were observed when action potentials were blocked by TTX, these changes in frequencies of IPSCs and EPSCs recorded in the absence of TTX must represent differences in action potential-dependent neurotransmission. Therefore, these results indicate that inhibitory synaptic input was decreased because of reduced firing frequency of GABAergic interneurons caused by *Scn1a* haploinsufficiency, whereas excitatory synaptic activity was increased as an indirect consequence of decreased inhibition.

Treatment of autism-related phenotypes in *Scn1a*$^{+/-}$ mice with clonazepam

Given that the autism-related phenotype and spatial-learning deficit in *Scn1a*$^{+/-}$ mice emerge from decreased Na$_V$1.1 activity in GABAergic interneurons, we reasoned that they could be rescued by increasing the strength of GABAergic transmission. To test this idea, we treated *Scn1a*$^{+/-}$ and wild-type mice with the benzodiazepine clonazepam, a positive allosteric modulator of the GABA$_A$ receptor. Benzodiazepines do not open the GABA$_A$ receptor chloride channel in the absence of GABA, but instead boost GABA signalling only when presynaptically released GABA binds to the receptor[39]. First, we examined the effects of clonazepam in the open-field and elevated plus-maze tests to avoid potential sedative and anxiolytic effects in our behavioural experiments, which depend on locomotor activity. The maximal intraperitoneal dose of clonazepam that did not cause significant sedation or anxiolytic effect in the open field and elevated plus maze tests was 0.0625 mg kg^{-1} for *Scn1a*$^{+/-}$ mice (Fig. 5a and Supplementary Fig. 21), 20-fold lower than typical anxiolytic doses[40]. To test the effect of clonazepam on social behaviour, we performed three sets of identical trials at one-week intervals with the same groups of mice. In the first trial, we performed the social interaction test in the open arena and the three-chamber test without any treatment. In a subsequent trial, the same behavioural tests were performed 30 min after intraperitoneal injection of 0.0625 mg kg^{-1} clonazepam. In the last trial, the tests were performed 30 min after intraperitoneal injection of vehicle. The data were analysed as the ratio of the time of interaction with a stranger mouse over the time of interaction with an empty cage. Both in the open arena and in the three-chamber test, clonazepam treatment completely rescued impaired social behaviours of the *Scn1a*$^{+/-}$ mice, and this effect was reversed after the one-week clearing period (Fig. 5b, c and Supplementary Figs 22 and 23). In contrast, low-dose clonazepam had no effect on the social behaviour of wild-type mice. Treatment with low-dose clonazepam 30 min before testing also rescued impaired context-dependent fear conditioning. Whereas wild-type mice were unaffected by clonazepam (Fig. 5d), *Scn1a*$^{+/-}$ mice showed a complete reversal of the loss of their 30-min and 24-h contextual fear memory (Fig. 5e). These results indicate that a single low dose of clonazepam can reversibly rescue core autistic traits and cognitive deficit in *Scn1a*$^{+/-}$ mice.

We also tested the effects of clonazepam on GABAergic inhibitory transmission in the hippocampal CA1 region in *Scn1a*$^{+/-}$ mice. As expected, treatment with 10 μM clonazepam increased sIPSC amplitude, but not frequency, in *Scn1a*$^{+/-}$ hippocampal slices (Fig. 5f and Supplementary Fig. 24a). The increased amplitude of spontaneous IPSCs after treatment with 10 μM clonazepam leads to a decrease in frequency of spontaneous EPSCs, without change in amplitude in *Scn1a*$^{+/-}$ hippocampal slices (Fig. 5g and Supplementary Fig. 24b). These results support our hypothesis that behavioural rescue by treatment with clonazepam is associated with increased strength of inhibitory transmission.

Discussion

Despite their adverse impacts on quality of life, the neuropsychiatric comorbidities and cognitive deficit in DS have not previously been studied in an animal model, and the role of the Na$_V$1.1 channel in these deficits in brain functions was unknown. Our results show that mice with heterozygous loss-of-function mutation in Na$_V$1.1 channels show both cognitive deficits and autistic traits, including hyperactivity, anxiety, excessive stereotyped behaviours and social interaction deficits. Together with previously reported phenotypes of epilepsy[9], premature death[9], thermally induced seizures[11], ataxia[10], and sleep dysfunction[12], these studies demonstrate that *Scn1a*$^{+/-}$ mice phenocopy all the major symptoms of DS.

These cognitive and behavioural deficits in *Scn1a*$^{+/-}$ mice are caused by decreased action potential firing in forebrain GABAergic interneurons. Our previous studies indicated that deletion of Na$_V$1.1 channels causes selective reduction in Na$^+$ currents and action potential firing of GABAergic interneurons in hippocampus and cerebellum[9,10]. This deficit in action potential firing in interneurons in the hippocampus leads to a selective loss of inhibitory neurotransmission compared to excitatory transmission (Fig. 4). Moreover, *Dlx1/2-Scn1a*$^{+/-}$ mice, which have a specific deficit in Na$_V$1.1

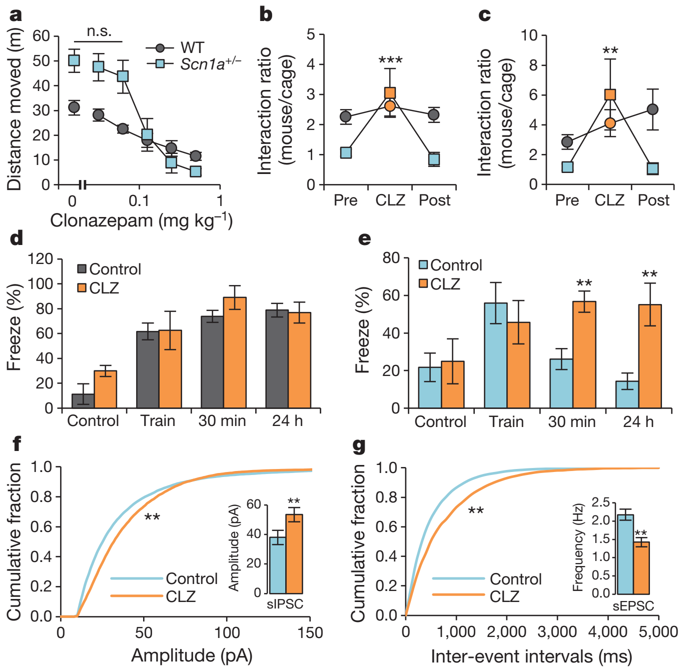


Figure 5 | Complete rescue of impaired social behaviour and fear-associated memory deficits by low-dose clonazepam treatment. **a**, Both wild-type and *Scn1a*$^{+/-}$ mice showed dose-dependent sedation by clonazepam (CLZ). Maximal concentration of CLZ without sedative or anxiolytic effect was 0.0625 mg kg^{-1}. **b**, **c**, In the open field social interaction test (**b**) and 3-chamber social preference test (**c**), decreased social interaction in *Scn1a*$^{+/-}$ mice was completely restored by a single intraperitoneal injection of 0.0625 mg kg^{-1} CLZ 30 min before the test. This CLZ effect on social interaction completely disappeared after 1 week of clearance in the same *Scn1a*$^{+/-}$ mice. CLZ effects on social interaction were absent in wild-type mice. Pre, pre clonazepam-treated; Post, post clonazepam-treated. **d**, **e**, In the contextual fear-conditioning test, a single intraperitoneal injection of 0.0625 mg kg^{-1} CLZ, 30 min before the training, led to a complete rescue of short-term (30 min) and long-term (24 h) fear-associated contextual memory in *Scn1a*$^{+/-}$ mice (**e**), but no significant change of fear-associated contextual memory by CLZ was observed in wild-type mice (**d**). All data shown are means ± s.e.m. from 6–12 mice per genotype. n.s., not significant. **f**, Cumulative plot and average value (inset) of sIPSC amplitude. Treatment with 10 μM CLZ increased the amplitude of sIPSC, but the frequency of sIPSC was unchanged by 10 μM CLZ in *Scn1a*$^{+/-}$ hippocampal CA1 slices (Supplementary Fig. 24a). **g**, Cumulative plot and average value (inset) of sEPSC frequency. Treatment with 10 μM CLZ decreased the frequency of sEPSC, but the amplitude of sEPSC was unchanged by 10 μM CLZ in *Scn1a*$^{+/-}$ hippocampal CA1 slices (Supplementary Fig. 24b). All data shown are means ± s.e.m. from 15–20 recordings per treatment group. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

channels in forebrain GABAergic interneurons, reproduce the core autistic features and cognitive deficits (Fig. 3). These results indicate that the autism-related traits in DS mice are caused by decreased inhibitory neurotransmission in GABAergic interneurons as a consequence of *Scn1a* haploinsufficiency.

To test this hypothesis further, we treated *Scn1a*$^{+/-}$ with clonazepam, a benzodiazepine, to reverse decreased GABAergic tone. High-dose benzodiazepine has been widely used to alleviate epileptic seizure[13] and anxiety-like behaviours[40], but not for rescuing major autism-related behaviours because of its sedative effects. Remarkably, a single low-dose clonazepam injection completely rescued deficits in social interactions and fear-associated contextual memory without sedative or anxiolytic effects in *Scn1a*$^{+/-}$ mice. The reversible rescue of cognitive deficit and autism-related behaviours by clonazepam at the time of training implies that these comorbidities in DS mice are not caused by recurrent seizure-induced excitotoxicity, but instead are caused by *Scn1a* haploinsufficiency and the resulting reduction of GABAergic transmission. These results indicate that low-dose benzodiazepine treatment could be a potential pharmacological intervention for cognitive deficit and autistic symptoms in DS patients.

Genome-wide association studies identified the chromosome 2q24.3 region, where the *SCN1A* gene is located, as an autism susceptibility locus[18,19]. Sequencing of the genomes of autistic patients identified mutations of *SCN1A* gene in familial autism[20]. Exome sequencing revealed that *de novo* mutations in the *SCN1A* gene cause autism[21]. Our results suggest the hypothesis that DS should be included in the category of ASD-related syndromes, such as fragile-X syndrome, Rett syndrome and Timothy syndrome[41]. With a prevalence of 1:20,000 births for DS and related *SCN1A* channelopathies[42], DS is less frequent than fragile-X syndrome (1:5,000) or Rett syndrome (1:10,000), but much more common than Timothy syndrome (<1:1,000,000). Interestingly, mutations in many ASD susceptibility genes also exhibit cytogenetic dysfunctions in GABAergic interneurons[24–29,43,44]. Thus, autistic traits in DS and in a broad range of ASDs may be caused by a reduction of GABAergic signalling. Our results suggest that low-dose benzodiazepine treatment may be effective in alleviating these autistic traits and cognitive impairment in DS and possibly in ASDs more broadly.

METHODS SUMMARY

Animals. The mice used for all behavioural analyses were 6–8-month-old adult male mice except D*lx*1/2 conditional mutant mice which were 3–5 months old. Adult mice 10 months old were used for immunohistochemical staining, and young mice 3–4 weeks old were used for electrophysiological recording. All behavioural tests were done blind to genotypes with age-matched littermate pairs of male mice. All experiments with animals were performed according to the National Institutes of Health Guide for Care and Use of Laboratory Animals and were approved by the University of Washington Institutional Animal Care and Use Committee.

Statistical analysis. All data are shown as mean ± s.e.m. and analysed using Student's *t*-test, one-way analysis of variance (ANOVA) with Tukey's post hoc comparison, and two-way ANOVA with Bonferroni's post hoc comparison. All the statistical analyses were done using Prism 4 (GraphPad). Details of particular tests in each experiment are described in the Supplementary Methods and full statistical tests and values for behavioural data are presented in Supplementary Table 1.

Received 4 January; accepted 27 June 2012.
Published online 22 August 2012.

1. Wolff, M., Casse-Perrot, C. & Dravet, C. Severe myoclonic epilepsy of infants (Dravet syndrome): natural history and neuropsychological findings. *Epilepsia* **47** (Suppl. 2), 45–48 (2006).
2. Genton, P., Velizarova, R. & Dravet, C. Dravet syndrome: the long-term outcome. *Epilepsia* **52** (Suppl 2), 44–49 (2011).
3. Li, B. M. *et al.* Autism in Dravet syndrome: prevalence, features, and relationship to the clinical characteristics of epilepsy and mental retardation. *Epilepsy Behav.* **21,** 291–295 (2011).
4. Brunklaus, A., Dorris, L. & Zuberi, S. M. Comorbidities and predictors of health-related quality of life in Dravet syndrome. *Epilepsia* **52,** 1476–1482 (2011).
5. Besag, F. M. Behavioral aspects of pediatric epilepsy syndromes. *Epilepsy Behav.* **5** (Suppl. 1), 3–13 (2004).
6. Mahoney, K. *et al.* Variable neurologic phenotype in a GEFS+ family with a novel mutation in SCN1A. *Seizure* **18,** 492–497 (2009).
7. Claes, L. *et al.* De novo mutations in the sodium-channel gene *SCN1A* cause severe myoclonic epilepsy of infancy. *Am. J. Hum. Genet.* **68,** 1327–1332 (2001).
8. Catterall, W. A. From ionic currents to molecular mechanisms: the structure and function of voltage-gated sodium channels. *Neuron* **26,** 13–25 (2000).
9. Yu, F. H. *et al.* Reduced sodium current in GABAergic interneurons in a mouse model of severe myoclonic epilepsy in infancy. *Nature Neurosci.* **9,** 1142–1149 (2006).
10. Kalume, F., Yu, F. H., Westenbroek, R. E., Scheuer, T. & Catterall, W. A. Reduced sodium current in Purkinje neurons from Na$_V$1.1 mutant mice: implications for ataxia in severe myoclonic epilepsy in infancy. *J. Neurosci.* **27,** 11065–11074 (2007).
11. Oakley, J. C., Kalume, F., Yu, F. H., Scheuer, T. & Catterall, W. A. Temperature- and age-dependent seizures in a mouse model of severe myoclonic epilepsy in infancy. *Proc. Natl Acad. Sci. USA* **106,** 3994–3999 (2009).
12. Han, S. *et al.* Na$_V$1.1 channels are critical for intercellular communication in the suprachiasmatic nucleus and for normal circadian rhythms. *Proc. Natl Acad. Sci. USA* **109,** E368–E377 (2012).
13. Catterall, W. A., Kalume, F. & Oakley, J. C. Na$_V$1.1 channels and epilepsy. *J. Physiol. (Lond.)* **588,** 1849–1859 (2010).

14. Westenbroek, R. E., Merrick, D. K. & Catterall, W. A. Differential subcellular localization of the R_I and R_{II} Na^+ channel subtypes in central neurons. *Neuron* **3,** 695–704 (1989).
15. Van Wart, A., Trimmer, J. S. & Matthews, G. Polarized distribution of ion channels within microdomains of the axon initial segment. *J. Comp. Neurol.* **500,** 339–352 (2007).
16. Ogiwara, I. *et al.* $Na_V1.1$ localizes to axons of parvalbumin-positive inhibitory interneurons: a circuit basis for epileptic seizures in mice carrying an *Scn1a* gene mutation. *J. Neurosci.* **27,** 5903–5914 (2007).
17. Cheah, C. S. *et al.* Specific deletion of $Na_V1.1$ channels in inhibitory interneurons causes seizures and premature death in a mouse model of dravet syndrome. *Proc. Natl Acad. Sci. USA*. (in the press).
18. Pescucci, C. *et al.* 2q24–q31 deletion: report of a case and review of the literature. *Eur. J. Med. Genet.* **50,** 21–32 (2007).
19. Ramoz, N., Cai, G., Reichert, J. G., Silverman, J. M. & Buxbaum, J. D. An analysis of candidate autism loci on chromosome 2q24–q33: evidence for association to the *STK39* gene. *Am. J. Med. Genet. B* **147B,** 1152–1158 (2008).
20. Weiss, L. A. *et al.* Sodium channels *SCN1A*, *SCN2A* and *SCN3A* in familial autism. *Mol. Psychiatry* **8,** 186–194 (2003).
21. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43,** 585–589 (2011).
22. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
23. Chao, H. T. *et al.* Dysfunction in GABA signalling mediates autism-like stereotypies and Rett syndrome phenotypes. *Nature* **468,** 263–269 (2010).
24. Paluszkiewicz, S. M., Martin, B. S. & Huntsman, M. M. Fragile X syndrome: the GABAergic system and circuit dysfunction. *Dev. Neurosci.* **33,** 349–364 (2011).
25. Peñagarikano, O. *et al.* Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits. *Cell* **147,** 235–246 (2011).
26. Hussman, J. P. Suppressed GABAergic inhibition as a common factor in suspected etiologies of autism. *J. Autism Dev. Disord.* **31,** 247–248 (2001).
27. Rubenstein, J. L. & Merzenich, M. M. Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes Brain Behav.* **2,** 255–267 (2003).
28. Markram, K. & Markram, H. The intense world theory – a unifying theory of the neurobiology of autism. *Front. Hum. Neurosci.* **4,** 224 (2010).
29. Yizhar, O. *et al.* Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477,** 171–178 (2011).
30. Moretti, P., Bouwknecht, J. A., Teague, R., Paylor, R. & Zoghbi, H. Y. Abnormalities of social interactions and home-cage behavior in a mouse model of Rett syndrome. *Hum. Mol. Genet.* **14,** 205–220 (2005).
31. Geschwind, D. H. Genetics of autism spectrum disorders. *Trends Cogn. Sci.* **15,** 409–416 (2011).
32. Stockhorst, U. & Pietrowsky, R. Olfactory perception, communication, and the nose-to-brain pathway. *Physiol. Behav.* **83,** 3–11 (2004).
33. Wang, Z. *et al.* Pheromone detection in male mice depends on signaling through the type 3 adenylyl cyclase in the main olfactory epithelium. *J. Neurosci.* **26,** 7375–7379 (2006).
34. Silverman, J. L., Yang, M., Lord, C. & Crawley, J. N. Behavioural phenotyping assays for mouse models of autism. *Nature Rev. Neurosci.* **11,** 490–502 (2010).
35. Yang, M. *et al.* Low sociability in BTBR T + tf/J mice is independent of partner strain. *Physiol. Behav.* http://dx.doi.org/10.1016/j.physbeh.2011.12.025 (8 January 2012).
36. Gomot, M. *et al.* Change detection in children with autism: an auditory event-related fMRI study. *Neuroimage* **29,** 475–484 (2006).
37. Long, J. E., Cobos, I., Potter, G. B. & Rubenstein, J. L. *Dlx1*&2 and *Mash1* transcription factors control MGE and CGE patterning and differentiation through parallel and overlapping pathways. *Cereb. Cortex* **19** (Suppl. 1), i96–i106 (2009).
38. Potter, G. B. *et al.* Generation of Cre-transgenic mice using *Dlx1/Dlx2* enhancers and their characterization in GABAergic interneurons. *Mol. Cell. Neurosci.* **40,** 167–186 (2009).
39. Rudolph, U. & Knoflach, F. Beyond classical benzodiazepines: novel therapeutic potential of $GABA_A$ receptor subtypes. *Nature Rev. Drug Discov.* **10,** 685–697 (2011).
40. Löw, K. *et al.* Molecular and neuronal substrate for the selective attenuation of anxiety. *Science* **290,** 131–134 (2000).
41. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nature Rev. Genet.* **9,** 341–355 (2008).
42. Yakoub, M., Dulac, O., Jambaque, I., Chiron, C. & Plouin, P. Early diagnosis of severe myoclonic epilepsy in infancy. *Brain Dev.* **14,** 299–303 (1992).
43. Rossignol, E. Genetics and function of neocortical GABAergic interneurons in neurodevelopmental disorders. *Neural Plast.* **2011,** 649325 (2011).
44. Paluszkiewicz, S. M., Olmos-Serrano, J. L., Corbin, J. G. & Huntsman, M. M. Impaired inhibitory control of cortical synchronization in fragile X syndrome. *J. Neurophysiol.* **106,** 2264–2272 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by Research Grants R01 NS25704 (W.A.C.), R01 MH075016 (H.O.d.) and R37 MH049428 (J.L.R.) from the National Institutes of Health and by a grant from the McKnight Foundation (W.A.C.). The authors thank E. Strakbein in the Machine Division at the University of Washington for making all the mazes for the behavioural experiments in this study.

Author Contributions W.A.C. and H.O.d. are co-senior authors. S.H., C.T., R.E.W., C.S.C., T.S., H.O.d. and W.A.C. designed the experiments. S.H., C.T., R.E.W., C.S.C. and T.S. performed the experiments. F.H.Y., C.S.C., G.B.P., J.L.R. and W.A.C. designed, prepared and characterized the genetically modified mouse lines. S.H., C.T., J.L.R., H.O.d. and W.A.C. wrote and revised the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to W.A.C. (wcatt@uw.edu) or H.O.d. (horaciod@uw.edu).

ARTICLE

doi:10.1038/nature11405

An anatomically comprehensive atlas of the adult human brain transcriptome

Michael J. Hawrylycz[1]*, Ed S. Lein[1]*, Angela L. Guillozet-Bongaarts[1], Elaine H. Shen[1], Lydia Ng[1], Jeremy A. Miller[1], Louie N. van de Lagemaat[2], Kimberly A. Smith[1], Amanda Ebbert[1], Zackery L. Riley[1], Chris Abajian[1], Christian F. Beckmann[3], Amy Bernard[1], Darren Bertagnolli[1], Andrew F. Boe[1], Preston M. Cartagena[4], M. Mallar Chakravarty[1,5], Mike Chapin[1], Jimmy Chong[1], Rachel A. Dalley[1], Barry David Daly[6], Chinh Dang[1], Suvro Datta[1], Nick Dee[1], Tim A. Dolbeare[1], Vance Faber[1], David Feng[1], David R. Fowler[7], Jeff Goldy[1], Benjamin W. Gregor[1], Zeb Haradon[1], David R. Haynor[8], John G. Hohmann[1], Steve Horvath[9], Robert E. Howard[1], Andreas Jeromin[10], Jayson M. Jochim[1], Marty Kinnunen[1], Christopher Lau[1], Evan T. Lazarz[1], Changkyu Lee[1], Tracy A. Lemon[1], Ling Li[11], Yang Li[1], John A. Morris[1], Caroline C. Overly[1], Patrick D. Parker[1], Sheana E. Parry[1], Melissa Reding[1], Joshua J. Royall[1], Jay Schulkin[12], Pedro Adolfo Sequeira[13], Clifford R. Slaughterbeck[1], Simon C. Smith[14], Andy J. Sodt[1], Susan M. Sunkin[1], Beryl E. Swanson[1], Marquis P. Vawter[13], Derric Williams[1], Paul Wohnoutka[1], H. Ronald Zielke[15], Daniel H. Geschwind[16], Patrick R. Hof[17], Stephen M. Smith[18], Christof Koch[1,19], Seth G. N. Grant[2] & Allan R. Jones[1]

Neuroanatomically precise, genome-wide maps of transcript distributions are critical resources to complement genomic sequence data and to correlate functional and genetic brain architecture. Here we describe the generation and analysis of a transcriptional atlas of the adult human brain, comprising extensive histological analysis and comprehensive microarray profiling of ~900 neuroanatomically precise subdivisions in two individuals. Transcriptional regulation varies enormously by anatomical location, with different regions and their constituent cell types displaying robust molecular signatures that are highly conserved between individuals. Analysis of differential gene expression and gene co-expression relationships demonstrates that brain-wide variation strongly reflects the distributions of major cell classes such as neurons, oligodendrocytes, astrocytes and microglia. Local neighbourhood relationships between fine anatomical subdivisions are associated with discrete neuronal subtypes and genes involved with synaptic transmission. The neocortex displays a relatively homogeneous transcriptional pattern, but with distinct features associated selectively with primary sensorimotor cortices and with enriched frontal lobe expression. Notably, the spatial topography of the neocortex is strongly reflected in its molecular topography—the closer two cortical regions, the more similar their transcriptomes. This freely accessible online data resource forms a high-resolution transcriptional baseline for neurogenetic studies of normal and abnormal human brain function.

The enormous complexity of the human brain is a function of its precise circuitry, its structural and cellular diversity, and, ultimately, the regulation of its underlying transcriptome. In rodents, brain- and transcriptome-wide, cellular-resolution maps of transcript distributions are widely useful resources to complement genomic sequence data[1–3]. However, owing to the challenges of a 1,000-fold increase in size from mouse to human, limitations in post-mortem tissue availability and quality, and the destructive nature of molecular assays, there has been no human counterpart so far. Several important recent studies have begun to analyse transcriptional dynamics during human brain development[4,5], although only in a small number of relatively coarse brain regions. Characterizing the complete transcriptional architecture of the human brain will provide important information for understanding the impact of genetic disorders on different brain regions and functional circuits. Furthermore, conservation and divergence in brain function between humans and other species provide essential information for the understanding of drug action, which is often poorly conserved across species[6].

The goal of the Allen Human Brain Atlas is to create a comprehensive map of transcript usage across the entire adult brain, with the emphasis on anatomically complete coverage at a fine nuclear resolution in a small number of high-quality, clinically unremarkable brains profiled with DNA microarrays for quantitative gene-level transcriptome coverage. Furthermore, structural brain imaging data were obtained from each individual to visualize gene expression data in its native three-dimensional anatomical coordinate space, and to allow correlations between imaging and transcriptome modalities. These data are freely accessible via the Allen Brain Atlas data portal (http://www.brain-map.org).

[1]Allen Institute for Brain Science, Seattle, Washington 98103, USA. [2]Genes to Cognition Programme, Edinburgh University, Edinburgh EH16 4SB, UK. [3]MIRA Institute, University of Twente & Donders Institute, Radboud University Nijmegen, Nijmegen, Netherlands. [4]Department of Psychiatry & Human Behavior, University of California, Irvine, California 92697, USA. [5]Kimel Family Translational Imaging-Genetics Laboratory, Centre for Addiction and Mental Health Toronto, Ontario M5S 2S1, Canada. [6]University of Maryland School of Medicine, Department of Diagnostic Radiology, University of Maryland Medical Center, Baltimore, Maryland 21201, USA. [7]Department of Pathology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [8]Department of Radiology, University of Washington, Seattle, Washington 98195, USA. [9]Department of Human Genetics, Gonda Research Center, David Geffen School of Medicine, Los Angeles, California 90095, USA. [10]Banyan Biomarkers, Inc., Alachua, Florida 32615, USA. [11]Office of the Chief Medical Examiner, Baltimore, MD, Department of Pediatrics, University of Maryland, Baltimore, Maryland 21201, USA. [12]Department of Neuroscience, Georgetown University, School of Medicine, Washington DC 20007, USA. [13]Functional Genomics Laboratory, Department of Psychiatry & Human Behavior, School of Medicine, University of California, Irvine, California 92697, USA. [14]Histion LLC, Everett, Washington 98204, USA. [15]The Eunice Kennedy Shriver NICHD Brain and Tissue Bank for Developmental Disorders, University of Maryland, Baltimore, Maryland 21201, USA. [16]Program in Neurogenetics, Department of Neurology and Department of Human Genetics, and Semel Institute, David Geffen School of Medicine-UCLA, Los Angeles, California 90095, USA. [17]Fishberg Department of Neuroscience and Friedman Brain Institute, Mount Sinai School of Medicine, New York, New York 10029, USA. [18]FMRIB, Oxford University, Oxford OX3 9DU, UK. [19]Computation & Neural Systems, California Institute of Technology, Pasadena, California 91125, USA.

*These authors contributed equally to this work.

Global mapping of transcript distributions

A tissue processing and data collection pipeline was established to image the brain and subsequently dissect tissue samples from approximately 900 anatomically defined sites for RNA isolation and microarray analysis (Fig. 1 and Supplementary Methods 1). Two complete normal male brains were analysed from donors aged 24 and 39 years and are referred to here as Brain 1 and Brain 2 (Supplementary Table 1). Briefly, cooled brains underwent *in cranio* magnetic resonance imaging (MRI) followed by embedding, slabbing and freezing. Whole-brain cryosections were made from each slab, after which the slabs were subdivided and sectioned on 2 × 3 inch slides for histological analysis with Nissl and other markers for structure identification. Defined brain regions were isolated either using macrodissection (cortical gyri, other large structures) or laser microdissection (LMD; Leica LMD6000, Leica Microsystems) from tissue sections on polyethylene naphthalate (PEN) membrane slides (Leica Microsystems). Any given anatomical structure was first identified on the basis of histological data, and then sampled in a series of contiguous coronal slabs in both hemispheres. RNA was isolated from each sample and used to generate labelled cRNA probes for hybridization to custom 64K Agilent microarrays. The output of this pipeline was a set of microarrays that sample the entire spatial extent of neocortical gyri that could be reproducibly identified across individuals, as well as subcortical nuclear structures, at the resolution allowed by Nissl staining and sample size requirements for microarray analysis. One-hundred and seventy distinct structures were assayed at least once in both brains, and 146 structures twice or more (Supplementary Table 2). Sample locations were mapped back into the native brain MRI coordinates and subsequently to Montreal Neurological Institute (MNI) coordinate space[7].

These microarray data form the foundation for a publicly accessible online atlas, which includes viewers for microarray data visualization and mining, MRI/histology/sample location, and three-dimensional (3D) visualization of MRI and gene expression. To complement and validate the microarray data, several targeted, large-scale *in situ* hybridization (ISH) data sets were generated using a high-throughput ISH platform[1,8]. All of these data are linked with the other databases available via the Allen Brain Atlas data portal (http://www.brain-map.org) to facilitate comparative analyses with developing and adult mouse, rhesus macaque and human.

The output of the data generation pipeline described above is a detailed quantitative map of transcript distribution across the entire brain. As one example, Fig. 2a depicts the structural distribution of gene expression related to dopaminergic neurotransmission, illustrating the highly localized enrichment of genes associated with dopamine synthesis, packaging, degradation and postsynaptic signalling. Regional enrichments were conserved between the two brains (note similar peaks in paired rows for Brain 1 and 2; Fig. 2a) and were consistent with previous studies[9,10]. For example, tyrosine hydroxylase (*TH*) is enriched in the substantia nigra pars compacta (SNC), ventral tegmental area (VTA), and hypothalamic supraoptic and preoptic nuclei, as well as in the locus ceruleus, the neurons of which use dopamine as a precursor for noradrenaline. Similar brain-wide plots for other neurotransmitter systems are provided in Supplementary Fig. 1.

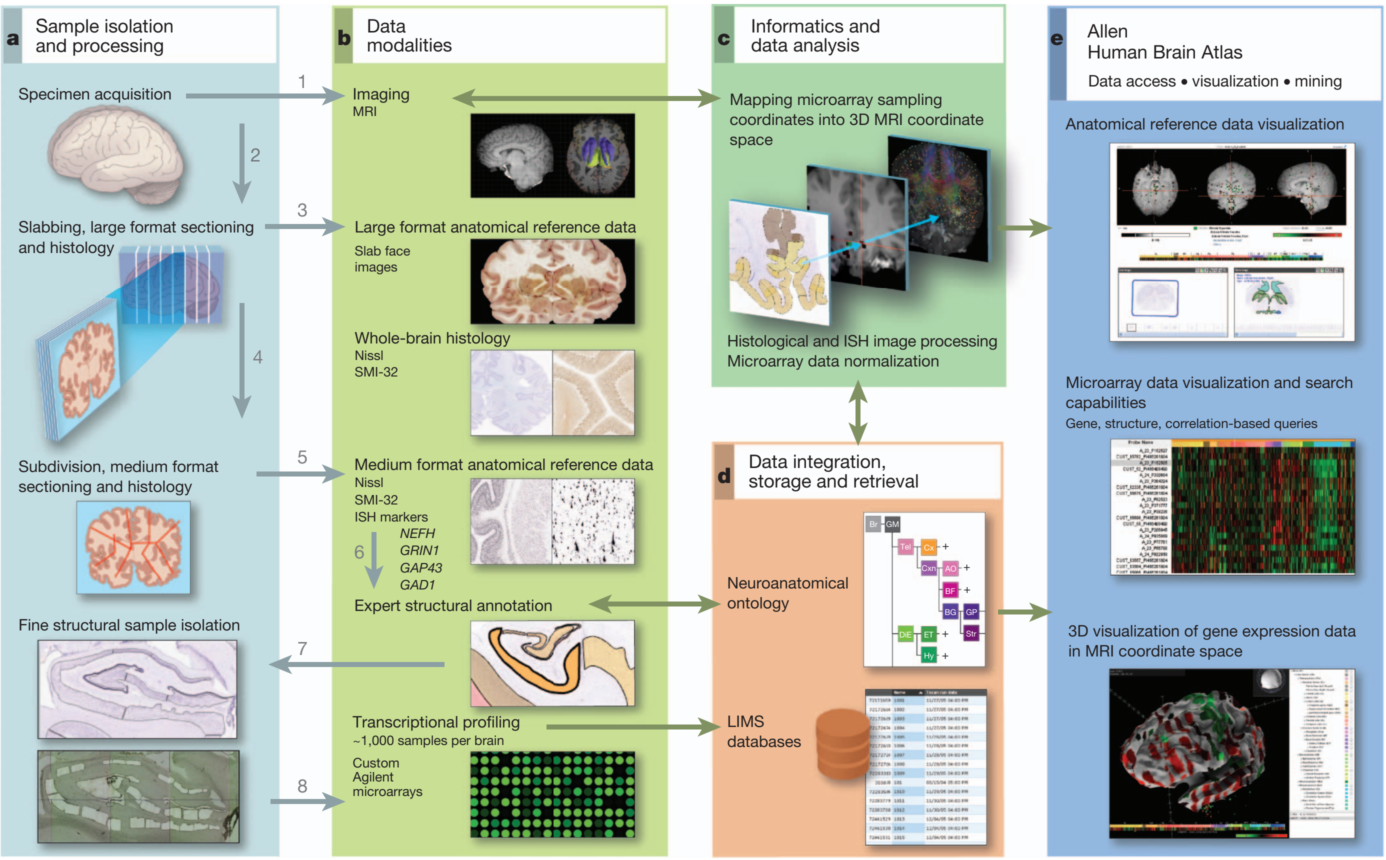

a
Sample isolation and processing
Specimen acquisition
Slabbing, large format sectioning and histology
Subdivision, medium format sectioning and histology
Fine structural sample isolation
1
2
3
4
5
6
7
8
b
Data modalities
Imaging
MRI
Large format anatomical reference data
Slab face images
Whole-brain histology
Nissl
SMI-32
Medium format anatomical reference data
Nissl
SMI-32
ISH markers
NEFH
GRIN1
GAP43
GAD1
Expert structural annotation
Transcriptional profiling
~1,000 samples per brain
Custom Agilent microarrays
c
Informatics and data analysis
Mapping microarray sampling coordinates into 3D MRI coordinate space
Histological and ISH image processing
Microarray data normalization
d
Data integration, storage and retrieval
Neuroanatomical ontology
Br
GM
Tel
Cx
Cxn
AO
BF
BG
GP
DiE
ET
Str
Hy
LIMS databases
e
Allen Human Brain Atlas
Data access • visualization • mining
Anatomical reference data visualization
Microarray data visualization and search capabilities
Gene, structure, correlation-based queries
3D visualization of gene expression data in MRI coordinate space

Figure 1 | Data generation and analysis pipeline. **a**, Experimental strategy to subdivide intact brains and isolate precise anatomical samples. **b**, Anatomical reference data are collected at each stage, including whole-brain MRI, large-format slab face and histology, medium (2 × 3-inch slide) format Nissl histology and ISH, and images of dissections. In Brain 2, labelling was performed for additional markers as shown. Histology data are used to identify structures, which are assembled into a database using a formal neuroanatomical ontology (**d**), and to guide laser microdissection of samples (**a**, lower panel). Isolated RNA is used for microarray profiling of ~900 samples per brain (**b**, lower panel). **c**, Microarray data are normalized and sample coordinates mapped to native 3D MRI coordinates. **e**, Data visualization and mining tools underlie the online public data resource. Numbers in **a** and **b** denote the order of sample processing steps leading to microarray data generation.

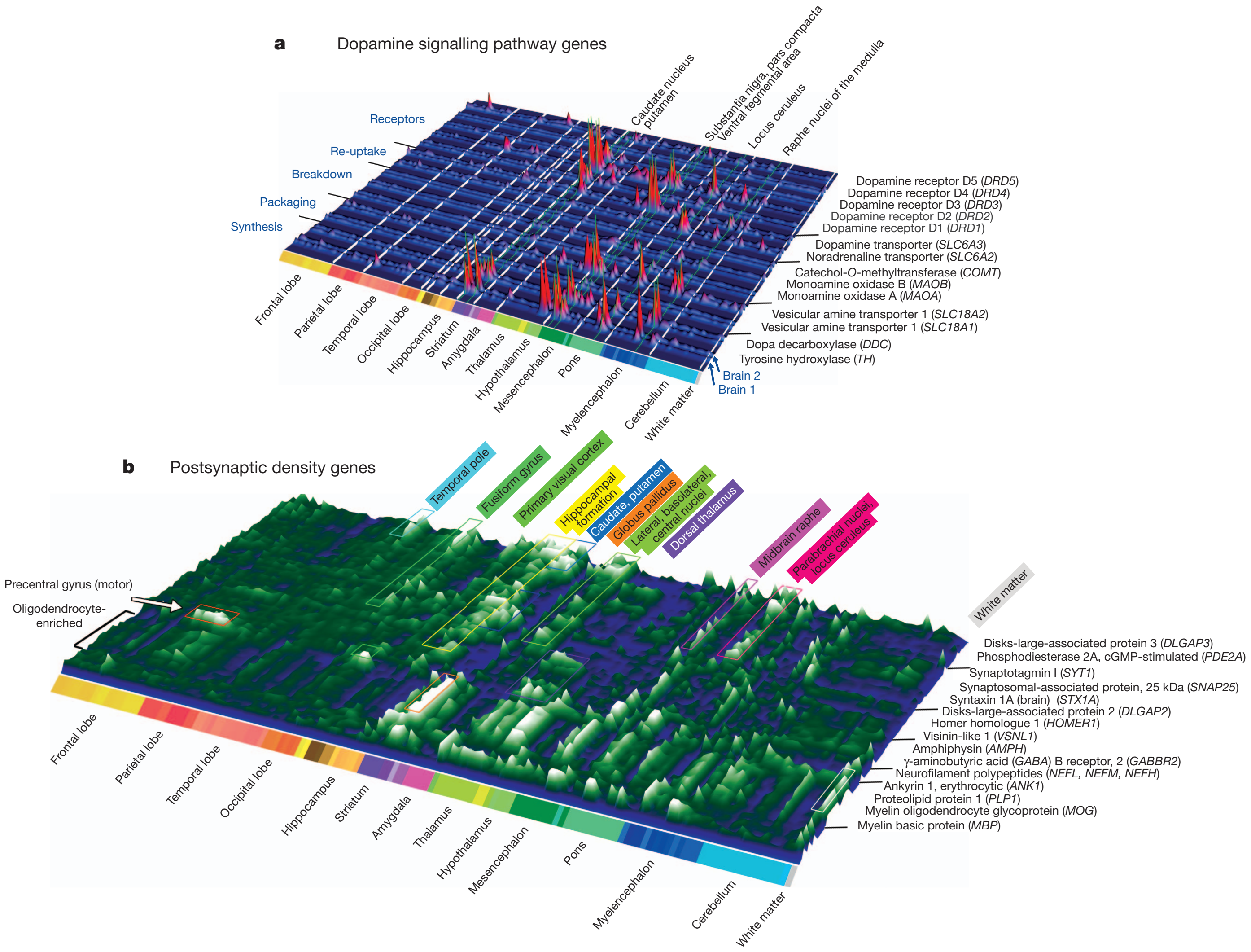


Figure 2 | Topography of transcript distributions for dopamine-signalling- and postsynaptic-density-associated genes. a, Gene expression profiles of genes associated with dopamine signalling plotted across 170 brain structures in two brains. Expression profiles for each probe plotted as raw microarray data normalized to mean structural expression, in paired rows to demonstrate consistency between the two brains. **b**, Gene-clustered topographic representation of the 74 most differentially expressed genes in human PSD preparations[12]. Gene profiles represent average expression in each structure between brains, plotted as deviation from the median. Clusters correspond to selective spatial enrichment of genes related to synaptic function, as well as an oligodendrocyte-enriched gene cluster (front cluster).

Interestingly, no statistically significant hemispheric differences could be identified at this fine structural level that were corroborated in both brains (paired one-sided *t*-tests, $P < 0.01$, Benjamini–Hochberg (BH)-corrected). Although surprising given well described lateralization of function, this finding is consistent with a recent study of developing human neocortex that failed to identify hemispheric differences despite extensive efforts using microarrays and quantitative PCR[11]. It may be that the basis for lateralization of function involves more subtle changes in specific cellular components, differences in relative area rather than type of functional domains between hemispheres, or is more related to functional connectivity patterns than molecular differentiation. Given this observation and to increase statistical power, samples from the two hemispheres for each structure were pooled for all subsequent analyses. In each brain independently, 84% of unique transcripts on the microarrays (29,412, referred to as genes for this manuscript) were found to be expressed in at least one structure (91.4% overlap in expressed gene sets between brains), consistent with the percentage of genes expressed in mouse brain by ISH (80%; ref. 1) and fetal human brain by microarrays (76%; ref. 11). Expression levels across anatomical structures were strongly correlated between brains (Pearson $r = 0.98$, $P < 10^{-40}$), with a highly significant correlation in differential expression relationships between structures (Pearson $r = 0.46$, $P < 10^{-40}$). Later in our analysis we completed data generation from a single (left) hemisphere of a third specimen. We found strong corroboration of overall expression levels and fold changes between structures in all three brains (Supplementary Fig. 2).

To illustrate the value of these data in understanding the functional organization of neurotransmission, we examined the 740 genes identified in the human excitatory postsynaptic density (PSD[12]), and in particular those that varied in their neuroanatomical distribution. Thirty-one per cent of PSD genes showed highly regional differential expression (Supplementary Methods 2 and Supplementary Table 3) (fold change >5 between any pair of 170 structures, false discovery rate <0.01), a significantly greater percentage than that observed across all genes (21%, $P < 10^{-6}$, Mann–Whitney *U*-test). As expected, many synapse-associated Gene Ontology (GO) categories[13] were enriched in this gene set, even relative to the PSD genes as a whole, including synapse (GO: 0045202), synaptic vesicle (GO: 0008021), synaptic transmission (GO: 0007268), neurophysiological process (GO: 0050877) and receptor activity (GO: 0004872).

Expression patterns for the most differentially expressed 10% of these PSD genes between any pair of structures are displayed in Fig. 2b

(74 genes with at least a 10.6-fold difference). The synapse-associated genes clustered into groups enriched in specific regions, indicative of a diverse set of excitatory synapse subtypes. For example, the primary motor cortex in the precentral gyrus, the origin of the longest range projection neurons, is delineated by selective enrichment of neurofilament proteins *NEFL*, *NEFM* and *NEFH*, which are frequently enriched in long-range projection neurons[14]. Surprisingly, a number of the most differentially expressed PSD-associated genes seem to be synthesized by glia, an observation made obvious by the stereotyped structural distribution of oligodendrocytes in white matter and other brain regions and the presence of well known myelin-associated genes (for example, myelin oligodendrocyte glycoprotein, *MOG*; myelin basic protein, *MBP*) in this gene cluster (Fig. 2b, front rows). The presence of these proteins in PSD preparations may represent a carry-over of glial fragments. Alternatively, they may be components of glutamatergic synapses between neurons and oligodendrocytes, which have been shown to share many properties with neuronal–neuronal synapses[15]. Overall, these data show remarkable regional variation in synaptic gene expression that probably underlies functional distinctions between regions.

Global transcriptional architecture of the adult brain

We next investigated the dominant features of transcriptional variation across the brain, beginning with global, brain-wide analyses and moving towards targeted local analyses of specific regions. An informative method for identifying biologically relevant patterns in high-dimensional microarray data sets is weighted gene co-expression network analysis (WGCNA)[16,17], which groups genes into modules that have strongly covarying patterns across the sample set. This method can identify gene expression patterns related to specific cell types such as neurons and glia from heterogeneous samples such as whole human cortex[18], due to the highly distinct transcriptional profiles of these cell types and variation in their relative proportions across samples. Each module is represented by an 'eigengene' corresponding to its expression pattern across structures (first left singular vector of the gene × structure matrix[16]), and genes highly correlated with the module eigengene are called 'hub' genes. This unbiased approach allows a module's function or cellular specificity to be imputed based on hub gene function, and allows statistical comparison either across studies to assign function or between brains to examine preservation between individuals.

Applied to the entire 911 sample set from Brain 1, genes were grouped into well-defined co-expression modules with specific anatomical distributions (Fig. 3a, b), consistent with previous studies in brain tissues[18,19]. Gene modules were frequently related to primary neural cell types and molecular functions (Fig. 3b, c). Several modules identify genes with enriched expression in neurons (M1–M2), based on overlap with neural-cell-type-enriched gene sets identified in previous studies[18] (second row in Fig. 3b). Genes in these modules are enriched in the neocortex (fifth row in Fig. 3b), and in particular cortical divisions as shown in eigengene plots (Fig. 3c). Hub genes and enriched GO terms for these modules are associated with neuronal structure and function and energy metabolism, as might be expected given the high metabolic demands of neurons (Supplementary Table 4). Other modules showed subcortical enrichment and correspond to expression in different types of glia (M8–M12), including microglia, astrocytes and oligodendrocytes. Additionally, one module with striking anatomical specificity for the paraventricular thalamus and central glial substance (asterisks in M5 eigengene histogram, Fig. 3c) corresponded to expression in the ventricular ependymal lining and choroid plexus. One highly regionalized neuron-related module (M6) was enriched in the striatum (the dopamine receptor *DRD1* in Fig. 2a is a hub gene). Thus, a major feature of the adult brain transcriptome profiled in this manner is the degree to which anatomical variation reflects the cellular make-up of different brain regions, both neuronal and non-neuronal.

The gene modules identified in Brain 1 were well conserved in Brain 2 as a whole (Fig. 3b), both at the level of regional gene expression patterns (third row) and as measured by a module preservation index (fourth row) using a summary *Z*-statistic as described previously[20]. Modules corresponding to broad neural cell types also showed highly significant preservation compared to a previous study using human brain samples (ref. 19 and data not shown).

We next took a more direct approach to examine relationships between regions of the brain based on dissimilarity of gene expression, by tabulating genes exhibiting highly differential expression between all pairs of regions. Significant pairwise differential relationships (BH-corrected $P < 0.01$) were independently recorded in each brain and a threshold set for at least a 2.8-fold ratio between structures (Supplementary Table 5). Figure 4a illustrates the resulting neuroanatomical molecular 'blueprint' common to both brains, by plotting the number of genes differentially expressed between each pair of structures based on the 11,414 genes passing these criteria in both brains (individual brain maps in Supplementary Fig. 3).

Many features of the brain transcriptome are apparent with this visualization. Remarkably few differences are seen at this fold change threshold across the neocortex (Fig. 4a, upper left) and cerebellum (lower right), reflecting their stereotyped repetitive cytoarchitecture. Exceptions to this relative cortical homogeneity include the postcentral gyrus (primary sensory cortex), temporal pole (area 38) and primary visual cortex (area 17). In contrast, complex differential relationships were observed between specific nuclei in subcortical structures. The globus pallidus and striatum have highly distinct profiles, as do several specific subcortical regions including the midbrain raphe, pontine nuclei and inferior olivary complex. The magnitude of differential expression between pairs of structures is also strongly correlated with the number of differentially expressed genes between these structures (Pearson $r = 0.62$, $P < 10^{-16}$; Supplementary Fig. 4).

Interestingly, a large percentage of these common differentially expressed transcripts (48%, or 5,500 probes) are poorly annotated, including probes not mapped to the human genome (HG19; http://genome.ucsc.edu/cgi-bin/hgGateway), mapped to contig sequences, or not mapped to known GENCODE genes[21]. Approximately 10% of these transcripts had very high correlation with the co-expression modules identified above (Pearson $r > 0.7$; Supplementary Table 6). For example, 38 transcripts demonstrated high correlation with the striatal module (M6), and 87 transcripts with the oligodendrocyte-associated module (M12; Fig. 3c), providing anatomical 'guilt-by-association' annotation of these genes of previously unknown function for selective roles in striatal and myelin function, respectively.

Most genes with high variation across brain regions are not selective for a single major brain region; rather, they are expressed in multiple regions and non-uniformly within these regions (Supplementary Fig. 5). This suggests that many genes may be quite pleiotropic with respect to brain function, and that local gene regulation in specific cytoarchitectural nuclei is the most important level of resolution. To summarize the complexity of structural variation and examine the extent to which major brain regions display local enrichment in specific fine cytoarchitectural divisions, we created a specificity index for each major region that measures enrichment in subdivisions of that region. This index, defined as the ratio of expression in one subdivision relative to the remaining subdivisions in that region (Supplementary Methods 3 and Supplementary Table 7), measures transcriptional diversity within regions. The results in Fig. 4b bear strong similarity to the plot in Fig. 4a, again with the neocortex and cerebellum displaying the least internal heterogeneity. In contrast, subcortical regions with many well-defined nuclei show the greatest local heterogeneity, including the myelencephalon, mesencephalon, pons, hippocampus and hypothalamus. It is also possible to identify genes with either brain-wide (global) or within structure (local) ubiquity (Supplementary Table 8). Not surprisingly, these gene sets are enriched for cellular

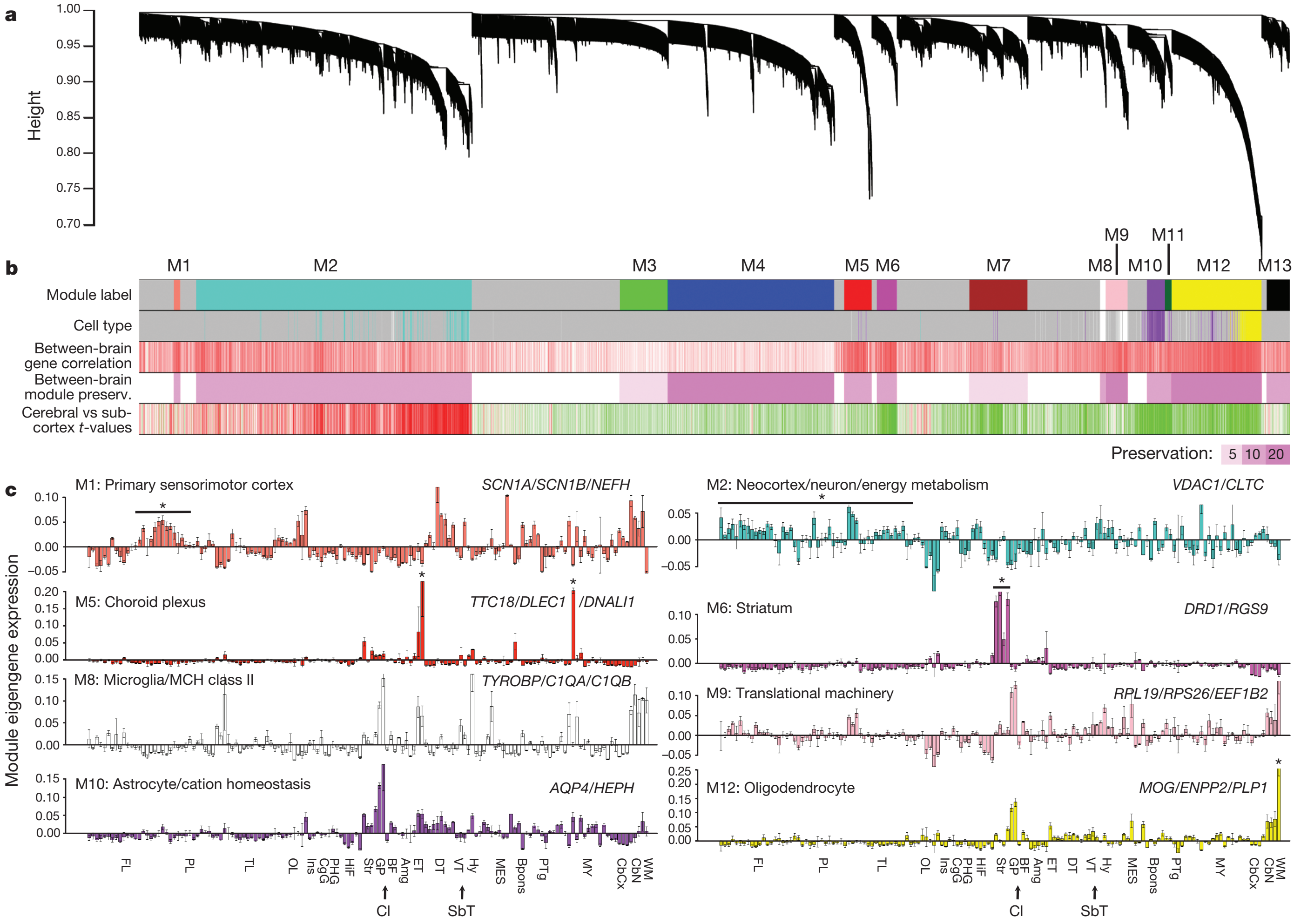


Figure 3 | Global gene networks. **a**, Cluster dendrogram groups genes into distinct modules using all samples in Brain 1, with the y axis corresponding to co-expression distance between genes and the x axis to genes (Supplementary Methods 2). **b**, Top colour band: colour-coded gene modules. Second band: genes enriched in different cell types (400 genes per cell type[18]) selectively overlap specific modules. Turquoise, neurons; yellow, oligodendrocytes; purple, astrocytes; white, microglia. Third band: correlation of expression across 170 subregions between the two brains. Red corresponds to positive correlations and white to no significant correlation. Fourth band: strong preservation of modules between Brain 1 and Brain 2, measured using a Z-score summary ($Z \geq 10$ indicates significant preservation). Fifth band: cortical (red) versus subcortical (green) enrichment (one-side t-test). **c**, Module eigengene expression (y axis) is shown for eight modules across 170 subregions with standard error. Dotted lines delineate major regions (see Supplementary Table 2 for structure abbreviations). An asterisk marks regions of interest. Module eigengene classifiers are based on structural expression pattern, putative cell type and significant GO terms. Selected hub genes are shown.

organelles and 'housekeeping' functions (for example, ribosome, mitochondrion, metabolism).

Local patterning reflects hippocampal cytoarchitecture

To explore local variation, we identified unique transcriptional signatures by analysis of variance (ANOVA) for the hippocampus. Following unsupervised hierarchical 2D clustering, cytoarchitecturally discrete subdivisions of the hippocampus (dentate gyrus, CA fields and subiculum) showed distinctive expression patterns sufficiently robust to cluster together like-samples while distinguishing subdivisions from one another (Fig. 5a). Interestingly, samples from the CA3 and CA4 subfields were not discriminable (intermixing in Fig. 5a), consistent with the view that CA4 is not a functionally distinct subfield from CA3 (ref. 22). Similarly robust regional clustering was observed in the mesencephalon, pons and myelencephalon (Supplementary Fig. 6 and Supplementary Table 9). Differential expression across hippocampal subfields could be validated by ISH. For example, the calcium-binding protein CALB1 has strong selectivity for the dentate gyrus relative to other hippocampal subdivisions in both brains (Fig. 5b), and cellular specificity for dentate gyrus granule neurons is demonstrated on an independent adult brain specimen by ISH in Fig. 5c. Hippocampal ISH data for CALB1 generated with the same histology platform in adult mouse[1] and rhesus macaque[23] allowed a phyletic comparison. Interestingly, expression in human differs from that in mouse (Fig. 5d) and rhesus monkey (Fig. 5e), where CALB1 is robustly expressed in CA1 and CA2 in addition to dentate gyrus.

Neocortical transcription reflects spatial topography

Our extensive neocortical sampling allowed us to investigate transcriptional variation across the neocortex in relation to spatial position and functional parcellation. Although highly differential expression between cortical regions is much less pronounced than between other brain regions (Fig. 4), many genes show statistically significant variation between lobes or gyri at a lower threshold. We first identified the 1,000 genes displaying the most significant variation in expression between 56 gyri in both brains (ANOVA, $P < 0.01$ BH-corrected, ranked by fold change between gyri; Supplementary Table 10). We then performed principal component analysis (PCA) on the 1,000 (genes) by 56 (sampled gyri) matrices for both brains. As shown in Fig. 6a–c, the first three principal components had striking selectivity for specific cortical regions (samples ordered by lobe and roughly rostral to caudal within each lobe) and were generally

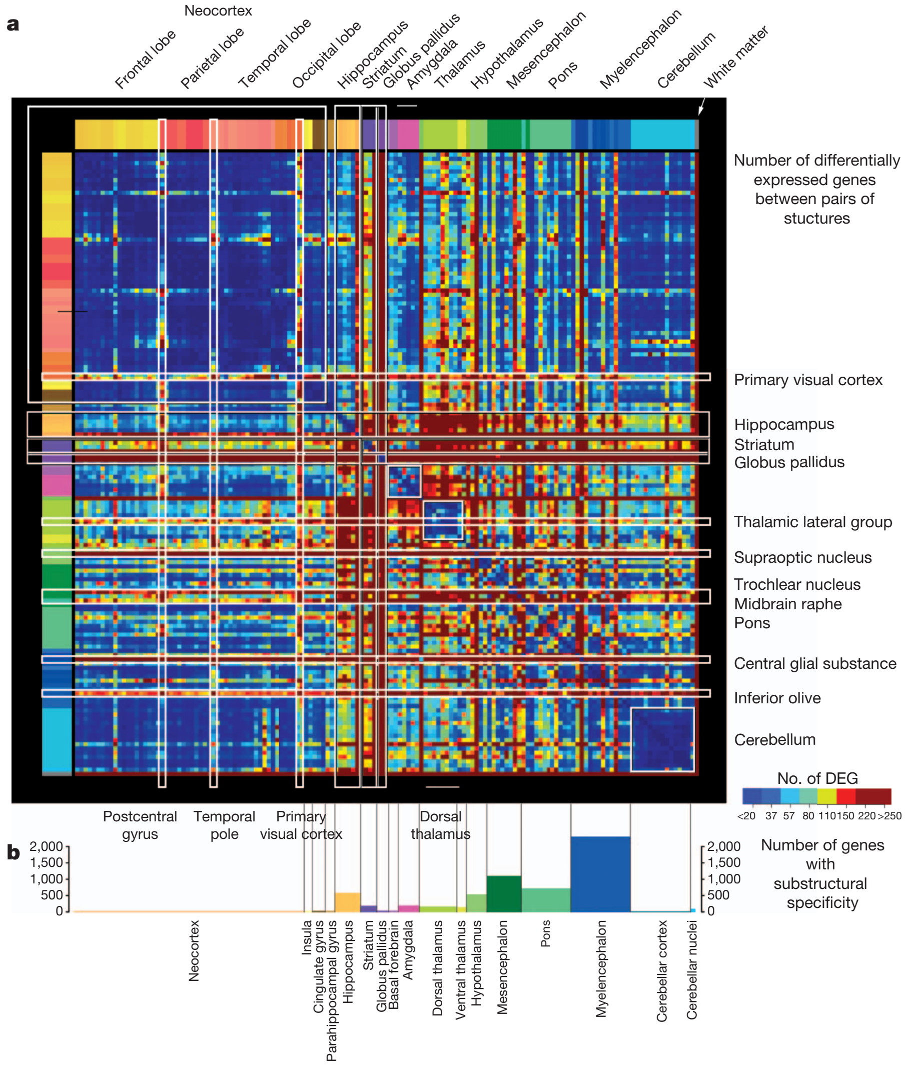


Figure 4 | Structural variation in gene expression. **a**, Matrix of differential expression between 146 regions in both brains. Each point represents the number of common genes enriched in one structure over another in both brains (BH-corrected $P < 0.01$, $\log_2$[fold change] > 1.5). DEG, differentially expressed genes. Several major regions exhibit relatively low internal variation (blue), including the neocortex, cerebellum, dorsal thalamus and amygdala. Subcortical regions show highly complex differential patterns between specific nuclei. **b**, Frequency of marker genes with selective expression in specific subdivisions of major brain regions (greater than twofold enrichment in a particular subdivision compared to the remaining subdivisions).

reproducible across both brains. PC1 is associated with primary sensorimotor cortices, with relative differential expression in precentral (motor) and postcentral (somatosensory) cortex, Heschl's gyrus (primary auditory) and primary and secondary visual areas. Confirmation of the visual cortex enrichment by ISH for several synaptic transmission-associated genes highly correlated to PC1 is shown in Supplementary Fig. 7. PC2 has areal selectivity for posterior orbital, paraolfactory and subcallosal gyri in the frontal lobe, the temporal pole, and the primary visual cortex. PC3 is primarily differential in frontal cortex compared to temporal and occipital cortex. These first three components accounted for a large amount of the variance (PC1: 58% in Brain 1, 42% in Brain 2; PC2: 10% in Brain 1, 11% in Brain 2; PC3: 5% in Brain 1, 8% in Brain 2; Supplementary Fig. 8). The spatial organization of the first three principal components was highly correlated between brains (Pearson $r = 0.71$ for PC1, 0.51 for PC2 and 0.70 for PC3).

To examine molecular relationships between different cortical regions, we applied multi-dimensional scaling (MDS; Supplementary Methods 3) to the samples of Brain 1 to visualize their genetic correlations along the directions of the first two (2D) or three (3D) principal components. Remarkably, the transcriptional relationships between samples recapitulate the spatial topography of the neocortex, as qualitatively illustrated after sample mapping in 2D (Fig. 6e). The relative positions of samples in the MDS plot mirror the actual positions of the gyri in the physical brain, shown in Fig. 6d on the MRI of the brain from which the samples were derived. Not only do samples from each lobe group together, but the relative positions of the lobes are anatomically correct. Furthermore, the relative position of each lobe's samples reflects the cortical topography, with the frontal pole and occipital striate cortex at opposite ends, precentral gyrus near postcentral gyrus, and so on. To provide a quantitative measure of this result we then applied the MDS method in 3D. As the positions of the samples were mapped back into MRI coordinate space, the correlation between 'genetic distance' and physical distance can be calculated after projecting the original cortical samples to a sphere and applying suitable rotation and scaling operations (Supplementary Methods 3). MDS-based sample correlations vary nearly linearly with 3D physical distance (Fig. 6e, inset), with a goodness of fit between native and MDS coordinates of 28.36% ($P < 10^{-4}$, Supplementary Fig. 9). This effect is strongest when limited to genes that are differential between gyri as above, but can also be seen using the entire ~30,000 gene set, achieving a fit of 12.48% between native and MDS coordinates ($P < 10^{-4}$, Supplementary Fig. 10). Therefore, gene expression profiles substantially determine position on the cortical sheet.

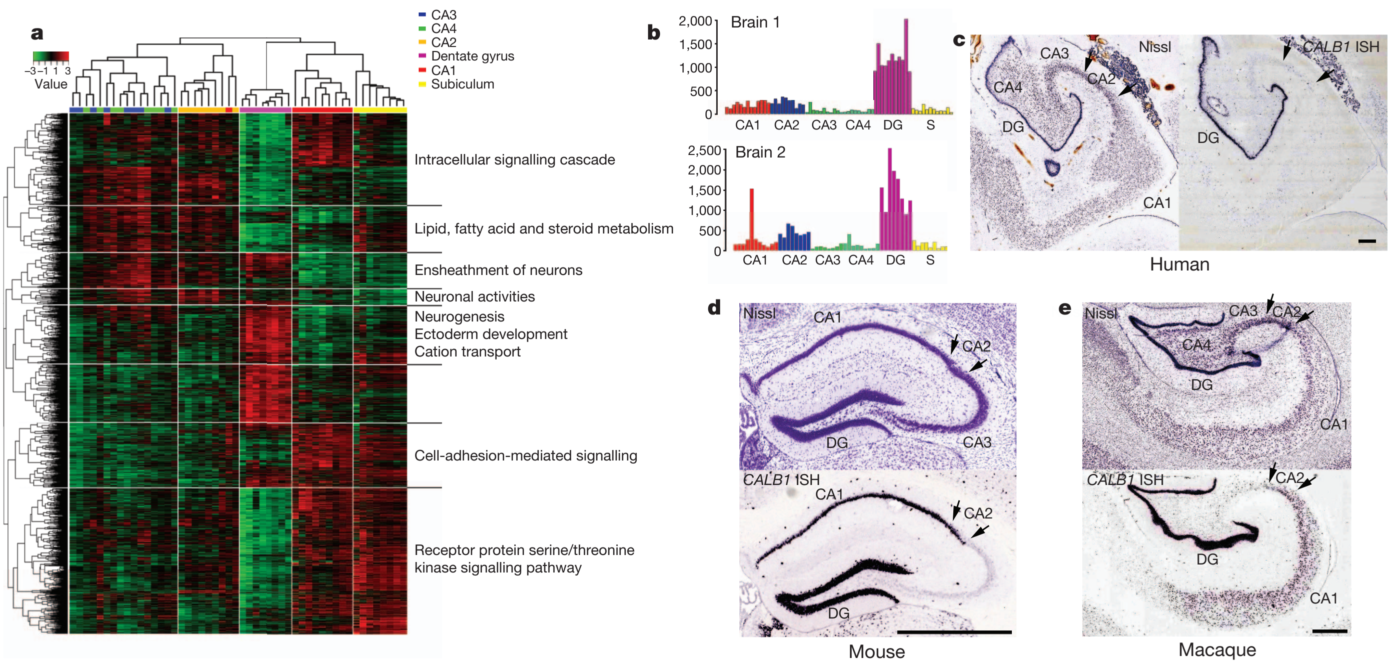


Figure 5 | Distinct transcriptional profiles of hippocampal subfields and human-specific pattern of *CALB1* expression. a, 2D clustering of microarray samples and differentially expressed genes across hippocampal subdivisions (ANOVA, $P < 0.01$ BH-corrected, top 5,000 genes), with selected enriched GO terms. **b**, Microarray data for *CALB1* shows enrichment in the dentate gyrus (DG) in both brains (*y* axis shows normalized raw microarray values). S, subiculum. **c**, Nissl (left) and *CALB1* ISH (right) through adult human hippocampus confirms dentate-gyrus-selective expression. **d**, **e**, Unlike human, *CALB1* ISH in the adult mouse (**d**) and rhesus macaque (**e**) show high *CALB1* expression in CA1 and CA2 (arrows) in addition to dentate gyrus. Scale bars: 1 mm.

Discussion

Molecular studies of human tissues are necessary for understanding the details of human brain function in the context of specific pathways and cell types and how they are affected in disease conditions. Here we describe the creation of an anatomically comprehensive transcriptional map in a small number of carefully selected, clinically unremarkable specimens, applying standardized digital molecular brain atlasing methods used in model organisms[3,24,25]. The combination of histology-guided fine neuroanatomical molecular profiling and mapping of gene expression data into MRI coordinate space produced an anatomically accurate quantitative map of transcript distribution across the entire human brain. This strategy was borne out in the robust differential molecular profiles of cytoarchitecturally and functionally distinct nuclei, providing a high-resolution genome-wide map of transcript distribution and the ability to analyse genes underlying the function of specific brain regions. Similar application of RNA sequencing methods[26,27], which were cost-prohibitive and technologically immature when the project was initiated, holds great promise for elucidating finer details of transcriptional regulation in the future.

Regional transcriptional signatures are highly conserved between the two brains assayed. These two individuals were males of similar age and ethnicity and therefore do not capture population or sex diversity; nevertheless, this high degree of similarity is suggestive of a strong underlying common blueprint for the human brain transcriptome and is consistent with other recent studies of human neocortical gene expression[4,5]. The availability of an entire hemisphere of a third brain specimen, as remarked above, enabled several confirmatory analyses to be performed. In particular, Supplementary Figs 11–13 report positively on the network analyses, structural variation of gene expression, and genetic topography of the neocortex. In summary, the high recapitulation of gene expression patterns across all three brains indicates that the basic transcriptional blueprint is robust across individuals. Ongoing work is focused on processing additional brains of both sexes to estimate the consistency of this blueprint.

The primary feature that distinguishes the human brain from that of other species is the enormous expansion of the neocortex relative to total brain volume. Our extensive profiling allowed us to ask directly how transcription varies across the neocortex. Surprisingly, we find a remarkable degree of transcriptional uniformity compared to other brain regions, apparently reflecting the similarity in laminar architecture across the entire neocortex[28]. However, there is significant, albeit less robust, variation in gene expression across cortical areas with two hallmark features. First, individual cortical samples showed such strong transcriptional similarities to neighbouring samples that the topography of the neocortex as a whole can, in part, be reconstructed based on their molecular profiles. One possible explanation is that these proximity relationships mirror lineage relationships of neocortical neurons generated from proximal parts of the developing neuroepithelium. Second, some primary sensory and motor regions do have distinct whole-transcriptome signatures, probably related to their specialized cellular and functional architecture. It is also likely that other more subtle features of cortical parcellation may not have been detected in the current analysis, including those identified using neurotransmitter receptor distributions[29] and functional connectivity[30]. One issue is that gyral patterns do not correlate perfectly with either cytoarchitectural or functional cortical parcellation. Greater regional differences may emerge if the samples can be grouped either by Brodmann area or on the basis of correlation to functional parcellations derived from functional imaging studies, now possible given the mapping of these data to MRI coordinates. Furthermore, it is likely that greater variation across areas will be found when assayed at the level of specific cortical cell types, as the excitatory neuron types in different layers display highly distinct molecular profiles[31] that have been shown to vary significantly across areas in primate neocortex[23]. Finally, higher confidence in consistent regional differences should emerge as more samples are investigated[32]. Nevertheless, the relative homogeneity of the two largest neuronal structures, with ~69 billion (cerebellar cortex) and ~16 billion (cortex) neurons out of the 86 billion neurons in the human brain[33], is striking and suggests an evolutionary expansion of a canonical cortical blueprint[34].

Finally, these data allow comparisons between humans and other animals, with particular relevance for studies of human disease. The

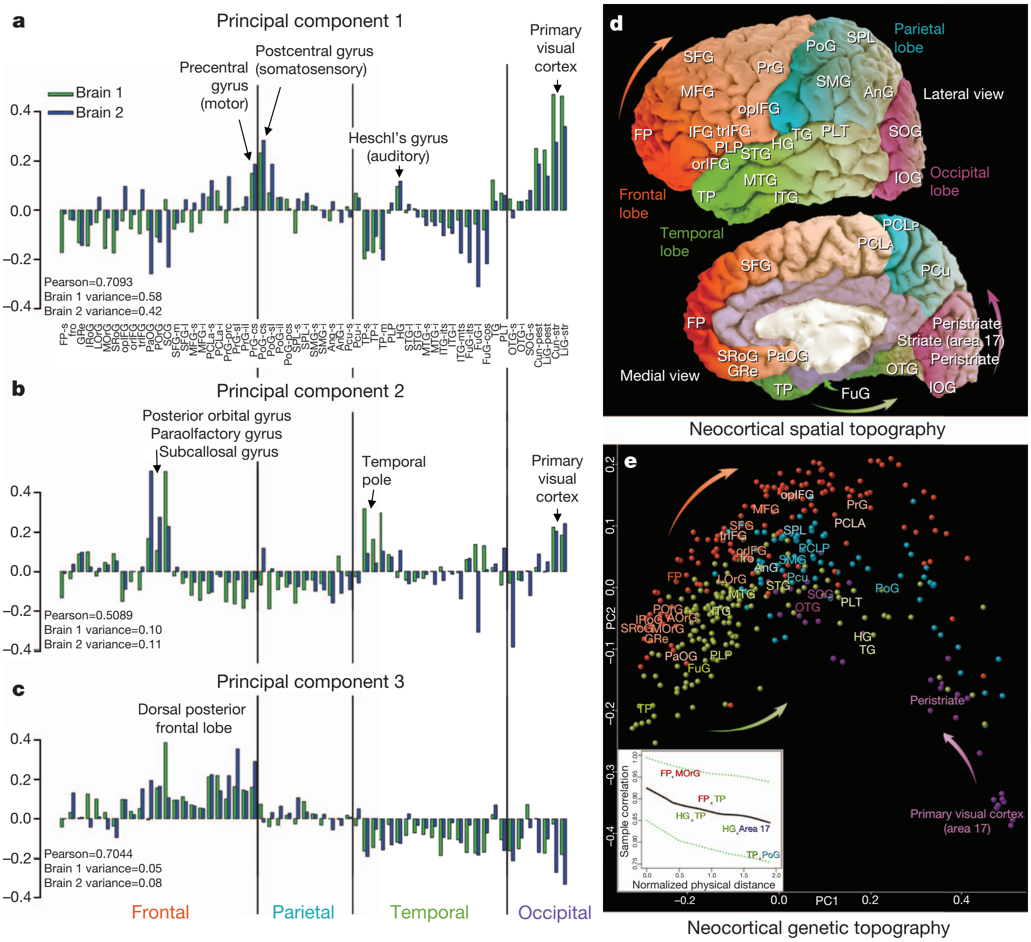


Figure 6 | The neocortical transcriptome reflects primary sensorimotor specialization and *in vivo* spatial topography. a–c, First three neocortical principal components, plotted across 57 cortical divisions ordered roughly rostral to caudal (frontal to occipital pole), are highly reproducible between brains. PC1 (Pearson $r = 0.71$) is selective for primary sensory and motor areas (**a**). PC2 (Pearson $r = 0.51$) is differential for specific subdivisions of the frontal, temporal and occipital poles (**b**), whereas PC3 (Pearson $r = 0.70$) is selective for the caudal portion of the frontal lobe (**c**). **d**, **e**, Relationship between the (x, y, z) location of sampled cortical gyri and their transcriptional similarities. Native Brain 1 MRI is shown in **d** with major gyri labelled (Supplementary Table 2). **e**, MDS applied to the same cortical samples, where distance between points reflects similarity in gene expression profiles. Median samples for major gyri are labelled. Samples cluster by lobe, and both lobe positions and gyral positions generally mirror the native spatial topography, emphasized by arrows in **d** and **e**. Inset panel in **e** plots the relationship (mean ± 1 s.d.) between 3D MDS-based similarity and 3D *in vivo* sample distance, demonstrating correlations that are stronger between proximal samples and decrease with distance. Selected gyral pairs are labelled. See Supplementary Table 2 for cortical gyrus abbreviations.

current manuscript describes a human-specific pattern for *CALB1* in the hippocampus compared to mouse and rhesus monkey. There are certain to be many such differences. In this light, these data should be extremely valuable from a translational perspective, allowing analysis of candidate genes and functional parcellation derived from genetic and imaging studies, and as a baseline for investigating neurological and neuropsychiatric disease.

METHODS SUMMARY

Anatomically comprehensive transcriptional profiling of adult human brains used high-throughput tissue processing and data generation pipelines for post-mortem brain imaging, anatomical delineation, sample isolation and microarray analysis. Data visualization and mining tools were developed to create a publicly accessible data resource (http://human.brain-map.org/). Extensive methodological details are supplied in Supplementary Methods 1.

Post-mortem tissue acquisition and screening. Tissue was provided by NICHD Brain and Tissue Bank for Developmental Disorders and the University of California, Irvine Psychiatry Brain Donor Program. After obtaining informed consent from decedent next-of-kin, specimens with no known neuropsychiatric or neuropathological history were collected and underwent serology, toxicology and neuropathological screening, and testing for RNA quality (RNA integrity number >6). Tissue collection was approved by Institutional Review Boards of the Maryland Department of Health and Hygiene, University of Maryland Baltimore and University of California Irvine. Specimens for microarray profiling were a 24-year-old African American male (Brain 1), a 39-year-old African American male (Brain 2), and a 57-year old Caucasian male (Brain 3; Supplementary Table 1).

Sample processing. Brains were imaged *in cranio* using MRI, cut into 0.5–1.0-cm-thick slabs and frozen. Slabs were subdivided and sectioned to allow histological staining, anatomical delineation and sample isolation using macro-dissection or laser microdissection. Total RNA was isolated and microarray data were generated by Beckman Coulter Genomics on Agilent 8 × 60K custom-design arrays (AMADID no. 024915). Sample locations were mapped from histology data into MR space using Inkscape (http://www.inkscape.org) and BioImage Suite (http://www.bioimagesuite.org) (Supplementary Methods 1).

Microarray data analysis. Weighted Gene Coexpression Analysis (WGCNA) was performed as described (Supplementary Methods 2)[16,17,20]. Module characterizations used Enrichment Analysis Systematic Explorer[35]. R (http://www.r-project.org/) was used for analysis and visualization (Supplementary Methods 2), principal component analysis (PCA), multidimensional scaling (MDS), and to transform MDS embedding into MNI space (Supplementary Methods 3).

***In situ* hybridization.** *In situ* hybridization used a semi-automated non-isotopic technology platform[1].

Received 22 December 2011; accepted 9 July 2012.

1. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445,** 168–176 (2007).
2. Diez-Roux, G. *et al.* A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.* **9,** e1000582 (2011).

3. Baldock, R. A. *et al.* EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics* **1,** 309–325 (2003).
4. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).
5. Colantuoni, C. *et al.* Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478,** 519–523 (2011).
6. Markou, A., Chiamulera, C., Geyer, M. A., Tricklebank, M. & Steckler, T. Removing obstacles in neuroscience drug discovery: the future path for animal models. *Neuropsychopharmacology* **34,** 74–89 (2009).
7. Evans, A. C. *et al.* Anatomical mapping of functional activation in stereotactic coordinate space. *Neuroimage* **1,** 43–53 (1992).
8. Zeng, H. *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149,** 483–496 (2012).
9. Bentivoglio, M. & Morelli, M. in *Handbook of Chemical Neuroanatomy. Dopamine* (eds Dunnett, S. B., Bentivoglio, M., Bjorklund, A. & Hokfelt, T.) Ch. I 1–107 (Elsevier, 2005).
10. Hurd, Y. L. & Hall, H. in *Handbook of Chemical Neuroanatomy. (Dopamine)* (eds S.B. Dunnett, M. Bentivoglio, A. Bjorklund, & T. Hokfelt) Ch. IX 525–571 (Elsevier, 2005).
11. Johnson, M. B. *et al.* Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62,** 494–509, doi:S0896–6273(09)00286–4 (2009).
12. Bayes, A. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature Neurosci.* **14,** 19–21 (2011).
13. Huang da, W. Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
14. Hof, P. R., Nimchinsky, E. A. & Morrison, J. H. Neurochemical phenotype of corticocortical connections in the macaque monkey: quantitative analysis of a subset of neurofilament protein-immunoreactive projection neurons in frontal, parietal, temporal, and cingulate cortices. *J. Comp. Neurol.* **362,** 109–133 (1995).
15. Bergles, D. E., Roberts, J. D., Somogyi, P. & Jahr, C. E. Glutamatergic synapses on oligodendrocyte precursor cells in the hippocampus. *Nature* **405,** 187–191 (2000).
16. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** (2005).
17. Horvath, S. *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl Acad. Sci. USA* **103,** 17402–17407 (2006).
18. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nature Neurosci.* **11,** 1271–1282 (2008).
19. Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl Acad. Sci. USA* **107,** 12698–12703 (2010).
20. Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7,** e1001057 (2011).
21. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (suppl. 1), 1–9 (2006).
22. Amaral, D. G. & Insausti, R. in *The Human Nervous System* (ed. Paxinos, G.) 771–755 (Academic, 1990).
23. Bernard, A. *et al.* Transcriptional architecture of the primate neocortex. *Neuron* **73,** 1083–1099 (2012).
24. Hawrylycz, M. *et al.* Digital atlasing and standardization in the mouse brain. *PLoS Comput. Biol.* **7,** e1001065 (2011).
25. Shattuck, D. W. *et al.* Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage* **39,** 1064–1080 (2008).
26. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
27. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Struct. Mol. Biol.* **18,** 1435–1440 (2011).
28. DeFelipe, J. & Jones, E. G. *Cajal on the Cerebral Cortex: an Annotated Translation of the Complete Writings* (Oxford Univ. Press, 1988).
29. Zilles, K. *et al.* Architectonics of the human cerebral cortex and transmitter receptor fingerprints: reconciling functional neuroanatomy and neurochemistry. *Eur. Neuropsychopharmacol.* **12,** 587–599 (2002).
30. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1,** 1–47 (1991).
31. Belgard, T. G. *et al.* A transcriptomic atlas of mouse neocortical layers. *Neuron* **71,** 605–616 (2011).
32. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).
33. Herculano-Houzel, S. The human brain in numbers: a linearly scaled-up primate brain. *Front. Hum. Neurosci.* **3,** 31 (2009).
34. Douglas, R. J. & Martin, K. A. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* **27,** 419–451 (2004).
35. Hosack, D. A., Dennis, G. Jr, Sherman, B. T., Lane, H. C. & Lempicki, R. A. Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4,** R70 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We wish to thank the Allen Institute founders, P. G. Allen and J. Allen, for their vision, encouragement, and support. We express our gratitude to past and present Allen Institute staff members R. Adams, K. Aiona, A. Alpisa, J. Arnold, C. Bennet, K. Brouner, S. Butler, E. Byrnes, S. Caldejon, J. Campiche, A. Carey, J. Chen, C. Copeland, C. Cuhaciyan, T. Desta, N. Dotson, S. Faber, T. Fliss, E. Fulfs, G. Gee, T. Gilbert, L. Gourley, G. Gu, J. Heilman, N. Ivanov, K. Keyser, A. Kriedberg, J. Laoenkue, F. Lee, S. Levine, L. Luong, N. Mastan, N. Mosqueda, E. Mott, N. Motz, D. Muzia, K. Ngo, A. Oldre, E. Olson, J. Parente, J. Phillips, L. Potekhina, T. Roberts, K. Roll, D. Rosen, M. Sarreal, S. Shapouri, N. Shapovalova, C. Simpson, D. Simpson, M. Smith, N. Stewart, K. Sweeney, A. Szafer, L. Velasquez, U. Wagley, W. Wakeman, C. White and B. Youngstrom for their technical assistance. We thank C. Long for mechanical engineering contract work. We thank R. Gullapalli, A. McMillan and R. Morales for post-mortem magnetic resonance imaging and radiology interpretation of MR data; J. Cottrell, M. Davis, R. Johnson, K. Moraniec, R. Vigorito, A. Weldon and the NICHD Brain and Tissue Bank for Developmental Disorders for tissue acquisition and processing; J. Davis for donor coordination; F. Mamdani, M. Martin, E. Moon, L. Morgan, B. Rollins and D. Walsh for tissue processing and psychological autopsy (DW); D. Patel for magnetic resonance imaging; and J. Sonnen for consultation on tissue microneuropathology. We also thank the External RNA Controls Consortium (ERCC), the US National Institute of Standards (NIST) and Technology, and M. Salit for access to ERCC transcripts during Phase V testing. We are grateful to Beckman Coulter Genomics (formerly Cogenics) and their staff P. Hurban, E. Lobenhofer, K. Phillips, A. Rouse and S. Beaver for microarray data generation and design of the custom Agilent array. We also wish to thank the Allen Human Brain Atlas Advisory Council members D. Geschwind, R. Gibbs, P. Hof, E. Jones, C. Koch, C. Saper, L. Swanson, A. Toga and D. Van Essen for their scientific guidance and dedication to the successful execution of this project. The project described was supported in part by Grant Numbers 1C76HF15069-01-00 and 1 1C76HF19619-01-00 from the Department of Health and Human Services Health Resources and Services Administration Awards and its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Department of Health and Human Services Health Resources and Services Administration Awards. S.G.N.G and L.V.L. were supported by the MRCD, Wellcome Trust and European Union Seventh Framework Programme under grants 241498 EUROSPIN, 242167 SynSys, and 241995 GENCODYS Projects.

Author Contributions A.R.J., A.L.G.-B., E.H.S. and K.A.S. contributed significantly to overall project design. A.L.G.-B., E.H.S., K.A.S., A.E. and P.W. managed the tissue and sample processing in the laboratory. D.B., A.F.B., R.A.D., J.G., B.W.G., R.E.H., M.K., T.A.L., P.D.P., S.E.P., M.R., J.J.R. and B.E.S. contributed to tissue and sample processing. E.H.S. and Z.L.R. contributed to establishing the tissue acquisition pipeline. P.M.C., B.D.D., D.R.F., L.L., P.A.S., M.P.V. and H.R.Z. contributed to tissue acquisition and MR imaging. Z.L.R., A.B., M.M.C., N.D., A.J., J.M.J., E.T.L., S.C.S. and P.R.H. contributed to protocol development. S.D., J.M.J., C.R.S. and D.W. provided engineering support. A.L.G.-B., R.A.D., P.D.P., J.G.H., J.A.Mo., J.J.R. and B.E.S. contributed to the neuroanatomical design and implementation. L.N. and C.D. managed the creation of the data pipeline, visualization and mining tools. L.N., C.D. and C.C.O. contributed to the overall online product concept. L.N., C.A., M.C., J.C., T.A.D., D.F., Z.H., C.La., Y.L. and A.J.S. contributed to the creation of the data pipeline, visualization and mining tools. M.J.H., E.S.L., J.A.Mi., D.H.G., L.N.L., C.F.B., S.M.Sm., S.G.N.G., A.L.G.-B., E.H.S., K.A.S., A.B., D.B., V.F., J.G., D.R.H., S.H., C.Le., J.S., S.M.Su., P.R.H. and C.K. contributed to data analysis and interpretation. A.R.J. supervised the overall project, and the manuscript was written by M.J.H. and E.S.L. with input from other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.J.H. (mikeh@alleninstitute.org).

doi:10.1038/nature11375

Structural plasticity and dynamic selectivity of acid-sensing ion channel–spider toxin complexes

Isabelle Baconguis[1] & Eric Gouaux[1,2]

Acid-sensing ion channels (ASICs) are voltage-independent, amiloride-sensitive channels involved in diverse physiological processes ranging from nociception to taste. Despite the importance of ASICs in physiology, we know little about the mechanism of channel activation. Here we show that psalmotoxin activates non-selective and Na^+-selective currents in chicken ASIC1a at pH 7.25 and 5.5, respectively. Crystal structures of ASIC1a–psalmotoxin complexes map the toxin binding site to the extracellular domain and show how toxin binding triggers an expansion of the extracellular vestibule and stabilization of the open channel pore. At pH 7.25 the pore is approximately 10 Å in diameter, whereas at pH 5.5 the pore is largely hydrophobic and elliptical in cross-section with dimensions of approximately 5 by 7 Å, consistent with a barrier mechanism for ion selectivity. These studies define mechanisms for activation of ASICs, illuminate the basis for dynamic ion selectivity and provide the blueprints for new therapeutic agents.

Acid-sensing ion channels (ASICs)[1], members of the epithelial sodium channel/degenerin (ENaC/DEG) superfamily of cation channels[2,3], open a transmembrane pore upon exposure to low pH[4]. Primarily found in the central and peripheral nervous systems[5–7], ASICs perform diverse physiological roles that include nociception[8,9], mechanosensation[8], synaptic plasticity, learning and memory[7], and fear conditioning[10]. The ASIC subfamily is encoded by four genes that give rise to seven isoforms[11], of which ASIC1a is permeable to Na^+ and Ca^{2+} and is involved in ischaemic neuronal injury[12,13]. The ENaC channel[3], found throughout the human body, is crucial to the regulation of blood pressure[14] and is directly involved in Liddle's syndrome[15] and pseudohypoaldosteronism[16].

ASICs and ENaCs are trimeric[17], voltage-independent and Na^+-selective ion channels sensitive to the classic ENaC blocker amiloride[1,3]. Whereas ASICs display a selectivity of Na^+:K^+ ranging from 3 to 30:1 and are inhibited by micromolar concentrations of amiloride, ENaCs harbour a preference for Na^+:K^+ of more than 100:1 and are blocked by nanomolar concentrations of amiloride[2,18]. For both ASICs and ENaCs, Li^+ permeability is similar to that of Na^+ and monovalent ions larger than K^+, such as Cs^+, are generally impermeable[19]. However, the 'peak' and 'sustained' or 'steady-state' ionic currents carried by ASICs display variable ion selectivity and blocker sensitivity[20–25], properties reminiscent of the dynamic ion selectivity of trimeric P2X receptors[26]. At present there is no understanding of how ASICs adopt Na^+-selective and non-selective conformations with differential sensitivity to the blocker amiloride.

Activation of the ion channel pore in ASICs is classically conditioned by drops in extracellular pH from about 7.5 to pH 4–6 (ref. 4) with the currents of ASIC1a showing rapid and nearly complete desensitization[27]. Psalmotoxin (PcTx1), classified as an inhibitor cystine knot toxin from a South American tarantula[28,29], acts potently on ASIC1a, increasing the channel's affinity for protons[30] and, contingent on the species and splice variant of the channel, acts as an agonist, eliciting steady-state current, or as an antagonist, diminishing ion channel activation[31,32]. The action of PcTx1 as an antagonist confers both analgesic[33] and neuroprotective[12] properties.

Here we report crystallographic and electrophysiological studies of the action of PcTx1 on chicken ASIC1a, showing the determinants of toxin binding, the mechanism by which toxin binding opens the ion channel, and the architecture of non-selective and Na^+-selective conformations of the ion channel pore. Our studies inform mechanisms of gating and permeation in ASICs and ENaC/DEG channels and lay a foundation for development of new molecules for modulation of ion channel activity.

Function and architecture of ASIC1a–PcTx1 complex

PcTx1 slows desensitization of ASIC and yields a substantial steady-state current when applied to ASIC1a–ASIC1b chimaeras, and thus we investigated whether PcTx1 stabilizes open channel states of chicken ASIC1a[31]. We generated a chicken ASIC1a construct for structural studies by removing 13 and 63 residues from the amino and carboxy termini, respectively, yielding a channel with wild-type-like electrophysiological properties (Δ13; Supplementary Figs 1 and 2). Application of a pH 5.5 solution to Δ13 gives rapidly activating and desensitizing inward current, whereas perfusion of a saturating PcTx1 solution at pH 7.25 elicits a current that activates and decays over a time scale of seconds to yield a steady-state current (Fig. 1a and Supplementary Fig. 3). Subsequent application of saturating PcTx1 at pH 5.5 further activates peak current and steady-state currents.

The Δ13–PcTx1 complex forms crystals at pH 7.25 (high pH, PDB ID 4FZ1) that belong to the *R*3 space group, diffract to approximately 3.3 Å resolution and have a single ASIC subunit–toxin complex positioned on the threefold axis of crystallographic symmetry. Crystals grown at pH 5.5 (low pH, PDB ID 4FZ0) belong to the *C*2 space group with an ASIC trimer and three toxin molecules in the asymmetric unit and a diffraction limit of about 2.8 Å resolution (Fig. 1b, c, d and

[1]Vollum Institute, Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA. [2]Howard Hughes Medical Institute, Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA.

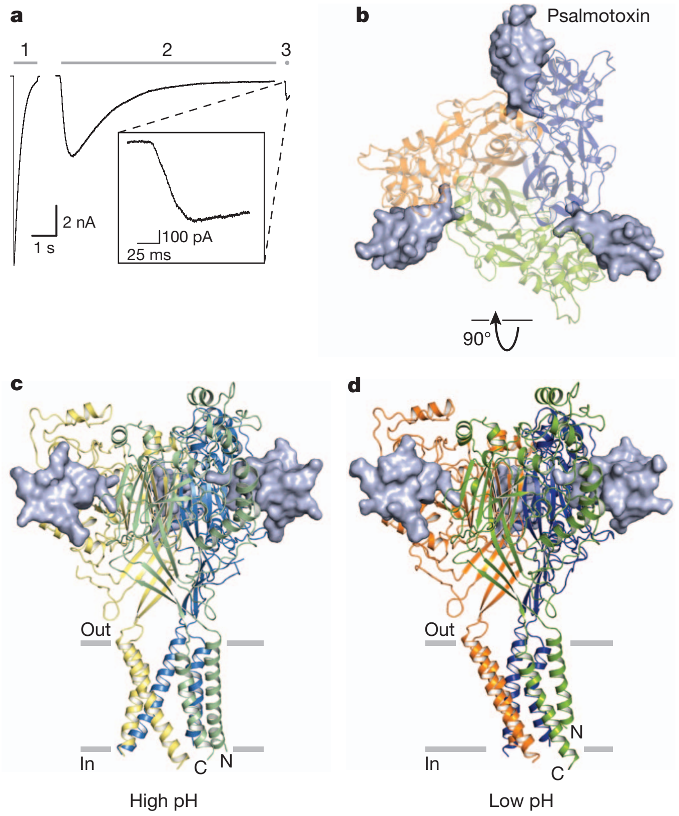


Figure 1 | PcTx1 activates the chicken ASIC1a Δ13 construct. a, Whole-cell, patch-clamp current traces of activation by steps into pH 5.5 (1), pH 7.25 and 1 μM PcTx1 (2), and pH 5.5 and 1 μM PcTx1 (3). Inset, current trace of step into pH 5.5 and 1 μM PcTx1. **b**, Structure of low-pH Δ13–PcTx1 complex viewed from the extracellular side. **c**, **d**, High-pH (**c**) and low-pH (**d**) complexes viewed parallel to the membrane. Each subunit is in a different colour and toxin is in solvent-accessible surface representation.

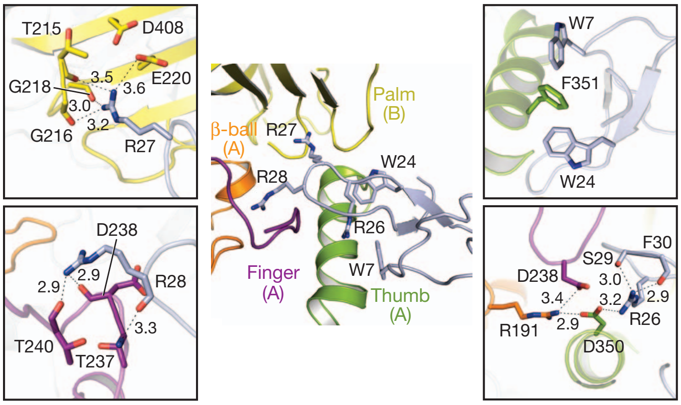


Figure 2 | Extensive interactions adhere PcTx1 to the Δ13 ion channel. Close-up views of toxin-binding site of the low-pH complex. Dashed lines indicate possible hydrogen bonds.

Supplementary Fig. 4). Structures of both crystal forms were solved by molecular replacement and refined to good crystallographic statistics (Supplementary Table 1, PDB IDs 4FZ0 and 4FZ1).

PcTx1 binds at subunit interfaces

The high- and low-pH channel–toxin complexes show three toxin molecules bound to the extracellular domain of each trimer at similar subunit interfaces approximately 45 Å from the transmembrane domain (Fig. 1). PcTx1 molecules bury approximately 900 Å of solvent-accessible surface area and make multiple ionic, polar and hydrophobic contacts, consistent with studies mapping sites of channel–toxin interaction[34] (Supplementary Fig. 5) yet distinct from computational modelling of the channel–toxin complex[35,36]. Docked on the thumb domain, toxin molecules form non-polar interactions mediated by aromatic residues, and they nestle an arginine-rich hairpin between adjacent subunits, making polar interactions in the acidic pocket (Fig. 2). Together, these interactions bridge the finger, β-ball and thumb domains of one subunit and the palm domain of the adjacent subunit. Arg 26 of PcTx1 tunnels under the toxin β-sheet to hydrogen-bond with the side chain carboxylate of Asp 350 of the thumb domain, which in turn interacts with Asp 238 and Arg 191 of the finger domain. In contrast, Arg 27 is oriented in the opposite direction, forming hydrogen bonds with backbone oxygens of Thr 215, Gly 216 and Gly 218, and also lying near the Glu 220–Asp 408 acidic pair, all on the palm domain of an adjacent subunit. Like Arg 27, Arg 28 interacts with backbone oxygens of Asp 238 and Thr 240 in the finger domain.

PcTx1 is further anchored on the thumb domain by an aromatic 'embrace' between Trp 7 and Trp 24 of PcTx1 and Phe 351 (Fig. 2), a residue that when mutated in human ASIC1a to leucine renders the channel insensitive to PcTx1 (ref. 37). Phe 351 is conserved in ASIC1 orthologues[17] and seems to be crucial to the specificity of PcTx1 to ASIC1 channels. Site-directed mutagenesis of PcTx1 (ref. 35) previously implicated Trp 24, Arg 26 and Arg 27 as central to toxin specificity, consistent with our structures of the channel–toxin complex. In addition, a recent structure of PcTx1 in complex with a non-functional construct of chicken ASIC1a supports our definition of the toxin-binding site[38].

Conformational changes in the extracellular domain

Structural comparisons of the high- and low-pH PcTx1-bound states with the desensitized state (PDB code 3HGC) show that the upper palm β-strands and knuckle domains define a structurally conserved scaffold (Supplementary Fig. 6), reminiscent of the scaffold of P2X receptors[39]. Relative to the desensitized state, the lower palm domain and the wrist of the high-pH Δ13–PcTx1 complex rotate by as much as 13° around an axis positioned below the scaffold (Fig. 3a). This rotation shifts subunit–subunit interfaces compared to those of the desensitized state structure, separating the thumb and palm domains of adjacent subunits by 2–3 Å and displacing the finger domains of the high- and low-pH complexes (Fig. 3a).

The consequences of toxin binding are manifested in the flexing of a blanket of β-sheets encapsulating the negatively charged central vestibule[40], a cavity composed of the lower palm domains, poised between the toxin binding site and the wrist (Fig. 3a and Supplementary Fig. 7). With the Cα atom of Val 75 as a landmark, the distances between adjacent subunits are approximately 11 Å and 12 Å in the low- and high-pH structures, respectively; upon formation of the desensitized state, the distance diminishes to about 7 Å. Chemical modification of the E79C mutant of ASIC3 (ref. 41), a residue equivalent to Glu 80 in chicken ASIC1a and predicted to face the interior of the central vestibule, slows the rate and extent of channel desensitization, consistent with the notion that the central vestibule contracts upon transition from the PcTx1-bound states to the desensitized state. Small molecules that activate[23] or potentiate the steady-state current of ASIC3 (refs 22, 42), respectively, bind within the central vestibule[23], or at the subunit interface near Ala 82, and stabilize the central vestibule in an expanded conformation, recalling the ATP-dependent expansion of the extracellular vestibule of P2X receptors[39].

A striking conformational change in the palm domain, common to both PcTx1 complexes and located within the β1–β2 linker, is a ∼180° flip of the Thr 84–Arg 85 peptide bond (Supplementary Fig. 8), inducing a shift of β1 away from the equivalent β-sheet of the adjacent subunit. In addition, Ala 413, Leu 414 and Asn 415 in the β11–β12 linker in the high-pH structure adopt conformations in which the side chains of Leu 414 and Asn 415 have effectively

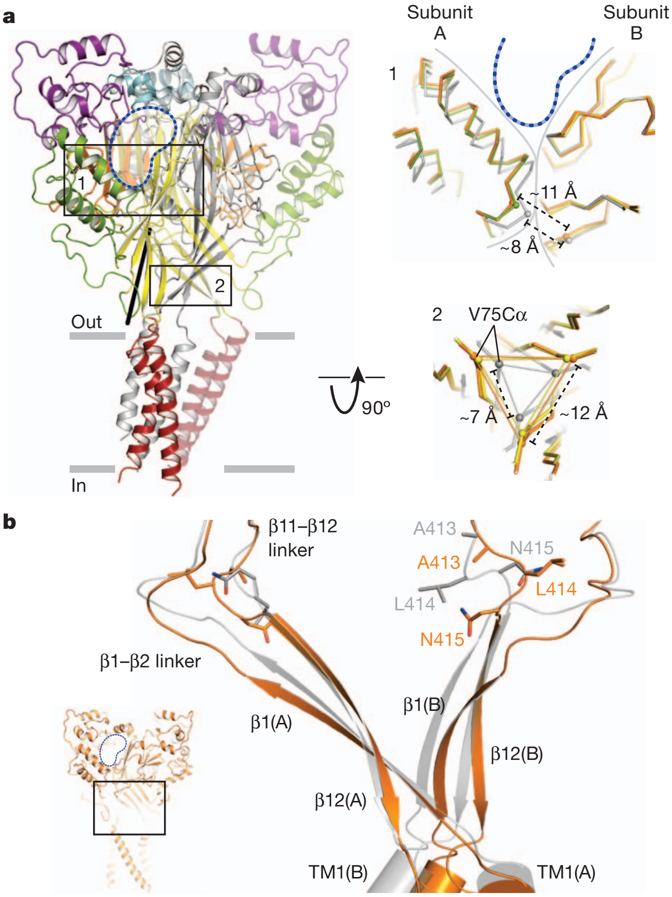


Figure 3 | Conformational changes in the extracellular domain. a, Low-pH Δ13–PcTx1 structure is in cartoon representation and coloured as in Fig. 2. Black line indicates the axis around which the lower palm domain and the wrist region rotate following superpositions of the desensitized and open state structures. Close-up views of selected regions are boxed. The low-pH complex is coloured by domain, the high-pH complex is orange, and the desensitized state (PDB code 3HGC) is grey. Approximate position of PcTx1 is indicated by blue dashed lines and the boundaries between adjacent subunits are shown by solid grey lines. Measured Cα distances are between residues Asn 357 (A) and Arg 85 (B) in (1). In (2) the distances are between Val 75 Cα atoms on adjacent subunits. **b**, Close-up view of strands β1 and β12, the β1–β2/β11–β12 linkers and the extracellular boundary of the transmembrane domains from two subunits of the high-pH Δ13–PcTx1 complex (orange) and the desensitized state (grey) following superposition of the respective scaffold domains. Inset shows location of close-up view in the context of the entire channel.

swapped positions (Fig. 3b and Supplementary Fig. 9). Indeed, the Cα atoms of residues Ala 82 and Ala 413 are farther apart in the high-pH state (8.1 Å) in comparison to the low-pH state (6.4 Å), consistent with the notion that a disulphide bond can form between the equivalent residues in shark ASIC1b, stabilizing the channel in the desensitized state[25] (Supplementary Fig. 10). Studies at sites equivalent to Leu 86, a residue that interacts with Leu 414 in the high-pH Δ13–PcTx1 state, and Asn 415 reinforce the conclusion that the β1–β2 and β11–β12 linkers are crucial to gating[42–44].

The structures of the high- and low-pH Δ13–PcTx1 complexes suggest molecular mechanisms underlying the pH-dependent conformations of the β1–β2 and β11–β12 linkers. We speculate that at high pH, the carboxyl group of Glu 80 is ionized, favouring an expanded conformation of the central cavity (Supplementary Fig. 7). As the pH drops, acidic residues in the vestibule bind protons, allowing Glu 80 and adjacent residues to adopt a ‘contracted’ central vestibule conformation. In fact, the distance between Cα of Glu 80 and Glu 412, residues that flank both linkers, diminishes from 14.1 Å to 11.3 Å in the high- and low-pH states, respectively. Neutralization of Glu 80 by mutation to Ala results in a channel that desensitizes approximately 40-fold more rapidly than the parent construct, and that does not yield measureable steady-state current upon application of PcTx1 at pH 7.25 (Supplementary Fig. 11), thus underscoring the role that titratable residues play in pH-dependent conformational changes of the central vestibule.

A highly conserved non-polar residue in the β11–β12 linker, Leu 414, is also important to the conformational transitions of the central vestibule. In the high-pH Δ13–PcTx1complex, Leu 414 interacts with Leu 86 of the β1–β2 linker. Upon transition to the low-pH Δ13–PcTx1 complex and to the desensitized state, however, Leu 414 forms multiple hydrophobic contacts involving Leu 281, Ile 306 and Val 368 of the adjacent subunit. In addition, Asn 415 hydrogen bonds with Tyr 416 and the backbone nitrogen of Ala 83 in the β1–β2 linker (Supplementary Fig. 9c, d). Together, these rearrangements highlight the importance of the β1–β2 and β11–β12 linkers and nearby subunit interfaces to conformational transitions of the central vestibule.

Ion channel at high pH

The high-pH Δ13–PcTx1 complex harbours a cavernous, threefold symmetric pore in which transmembrane helix 2 (TM2) resides on the periphery of the pore and lines the ion channel together with TM1 (Fig. 4a, b). In comparison to the desensitized state, a kink in TM2 at Asp 433 results in an effective rotation of the cytoplasmic end of TM2 by about 90° around the threefold axis and a tilting of TM2 by about 50°; movements in TM1 are smaller, characterized by an approximately 20° rotation around the threefold axis. Collectively, these movements rupture intra-subunit interactions between TM1 and TM2 and completely disrupt inter-subunit contacts between TM2 segments that define the closed, desensitized channel gate, yielding sparse inter-subunit and hydrophobic contacts between TM1 and TM2 mediated primarily by Ile 66 and Ile 434, and Val 46 and Ile 446 in the high-pH complex (Supplementary Fig. 12).

The large transmembrane pore of the pH 7.25 PcTx1 complex, with a diameter of about 10 Å near Asp 433 (Supplementary Fig. 13), led us to probe the selectivity of the Δ13–PcTx1 complex using bi-ionic patch-clamp electrophysiology (Fig. 4c, d). At neutral pH, the channel–toxin complex does not discriminate between Li^+, Na^+, K^+ and Cs^+, and we suggest that these ions permeate through the pore in a fully hydrated state. Furthermore, *N*-methyl-D-glucamine (NMDG), with a radius of about 4.0 Å, permeates through the ion channel, and application of 500 μM amiloride only blocks 10% of the steady-state current (Supplementary Fig. 14). This non-selective behaviour of the Δ13–PcTx1 complex is reminiscent of the non-selective behaviour previously observed in steady-state currents of wild-type ASIC3 and in the degenerin mutants of ASICs[20,21], and of the pore-dilated behaviour of trimeric P2X receptors[26].

Architecture of low-pH ion channel pore

The transmembrane helices of each subunit in the low-pH Δ13–PcTx1 complex not only adopt different conformations but also occupy distinct positions relative to the pore axis (Fig. 4e, f). TM2 segments of subunits A and B bend and ‘stretch’ at residues 433–435, similar to conformations observed in the high-pH structure. In comparison to the desensitized state, the TM1 helices of the low-pH complex rotate by 9–12° around the pore axis and the TM2 helices of subunits A, B and C tilt by approximately 47°, 65° and 30°, respectively. Thus the TM2 helices adopt an orientation nearly perpendicular to the membrane plane, reminiscent of the transmembrane conformation of the low-pH ΔASIC1 structure[17].

TM2 of the C subunit adopts a straight α-helix, resulting in an approximately four-residue displacement along the pore axis relative to subunits A and B, shifting subunit C towards the extracellular side of the membrane, thus conferring axial asymmetry onto the pore. This asymmetry has precedent in chemical modification studies of ENaC, which were interpreted in terms of a model in which the β subunit is

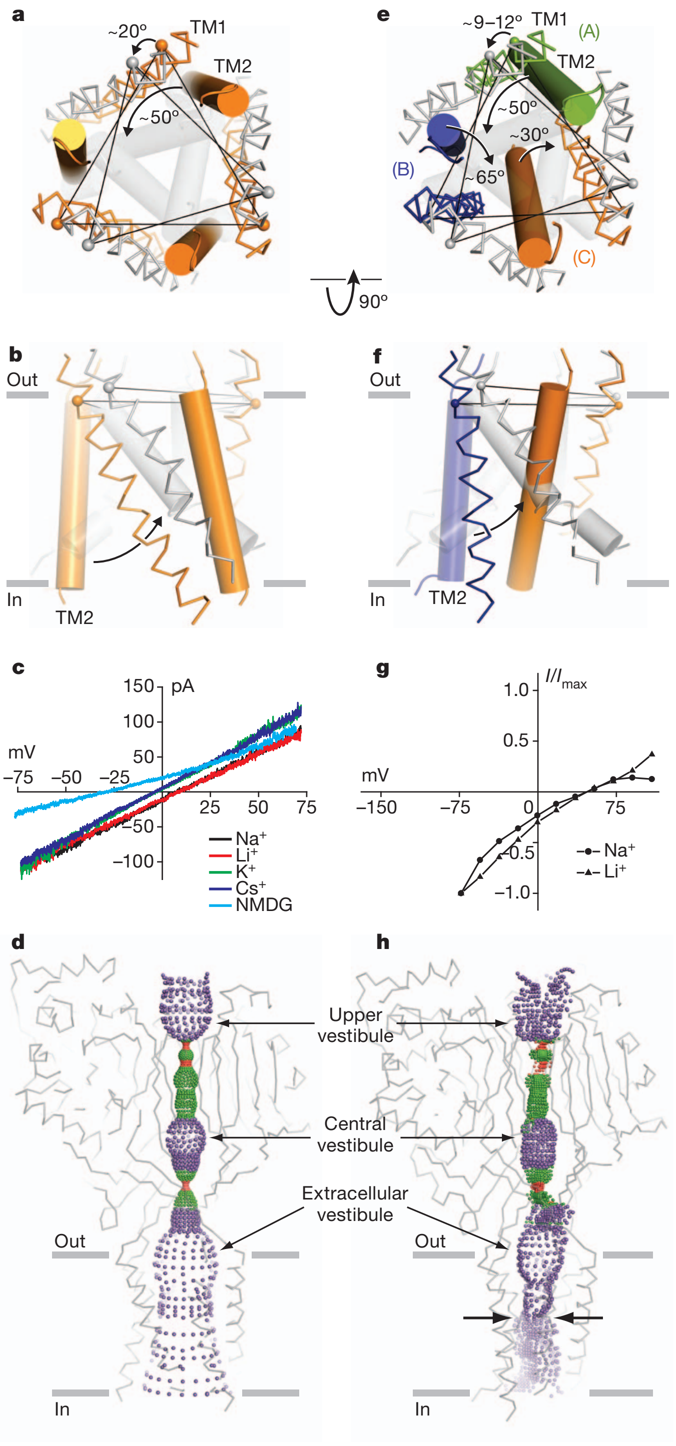


Figure 4 | Structural rearrangements and ion selectivity of the transmembrane pores. a, **b**, Comparison of transmembrane domains from the high-pH (orange) and desensitized (grey) state structures with TM1 in ribbon and TM2 in cylinder representation. Transmembrane domains are viewed from the extracellular side (**a**) and parallel to the membrane (**b**). **c**, Current/voltage experiment showing that at neutral pH the Δ13–PcTx1 complex forms a non-selective cation channel. **d**, Mapping of solvent-accessible pathway along the threefold axis shows that the high-pH complex has a large, transmembrane pore. The occluded pathway along the threefold axis in the extracellular domain indicates that ions access the pore by way of lateral fenestrations. **e**, **f**, A comparison of the transmembrane domains from the low-pH complex, where each subunit is in a different colour. The desensitized state is grey. TM1 and TM2 segments are in ribbon and cylinder representations, respectively. Transmembrane domains are viewed from the extracellular side (**e**) and parallel to the membrane (**f**). **g**, Current/voltage experiment demonstrating that the ion channel of the low-pH Δ13–PcTx1 complex is Na^{+}-selective. **h**, Mapping of a solvent accessible pathway along the pseudo threefold axis of the low-pH complex shows that it has an asymmetric ion channel pore and a constriction (opposing arrows) halfway across the bilayer. Maps of solvent-accessible pathways (**d** and **h**) were generated using the HOLE software (red < 1.4 Å < green < 2.3 Å < purple).

displaced along the pore axis by about one turn of an α-helix[45]. The axial displacement of the transmembrane segments gives rise to a striking constellation of hydrophobic contacts mediated by the staggered arrangement of Leu 440 and Leu 447 of subunits B and C that resembles a leucine zipper motif (Supplementary Fig. 15a, b). The extensive contacts between the TM2 domains of the B and C subunits give rise to TM1 of subunit B residing on the periphery of the ion channel domain, while the proximity of TM2 of subunits A and C shields TM1 of subunit C from exposure to the pore. The remaining four transmembrane segments line the pore and participate in intersubunit interactions that include contacts between Val 46 (subunit C) and Ile 446 (subunit A), as observed in the high-pH structure (Supplementary Fig. 15c, d). The exposure of portions of TM1 from subunit A to the pore concurs with the results from accessibility studies of TM1 residues of FaNaC[46], a peptide-gated channel, but stands in contrast to cysteine-directed chemical modification studies of lamprey ASIC[47], which suggest that TM2 primarily lines the pore. Determining how the conformation of the asymmetric pore in the low-pH Δ13–PcTx1 complex is related to the conductive pore of the Δ13 construct upon activation by protons will require additional studies. Nevertheless, we note that ENaCs may harbour an asymmetric pore on the basis of the variable reactivity of cysteine residues at equivalent sites on different subunits[45] and that the extensive intra- and intersubunit interactions between TM1 and TM2 seen in the low-pH Δ13–PcTx1 complex renders asymmetric pore formation favourable.

Low-pH pore is sodium-selective

The low-pH complex harbours an elliptical pore with a region of constriction spanning a helical turn and located approximately halfway along the transmembrane domain, primarily lined by hydrophobic side chains of Leu 440 (Supplementary Fig. 16). The position of the constriction exposes the putative amiloride binding site, Gly 439 and the 'GAS' selectivity tract, to the extracellular side of the membrane, consistent with the inhibitory mechanism of amiloride as an open channel blocker[48]. Notably, ion channel blockers of ASICs and ENaCs, including amiloride, are characteristically planar, non-symmetric molecules that roughly mirror the shape of the extracellular portion of the low-pH pore (Supplementary Fig. 16).

To assess the properties of the Δ13–PcTx1 ion channel pore at low pH, we determined its ion selectivity properties and found that the complex remains selective for Na^{+} and Li^{+} with a selectivity for Na^{+} over K^{+} of 10:1 (Fig. 4g). Because the pore is primarily lined by hydrophobic residues at the constriction point (Fig. 4h), we assume that Na^{+} ions are hydrated when traversing the pore. We suggest that the pore discriminates ions by the size of a fully or partially hydrated ion and that the mechanism underlying ion selectivity is best described by a barrier model. Indeed, it was proposed previously that Na^{+}-selective pores have a rectangular cross-section with dimensions of approximately 3.2 Å by 5.2 Å, large enough to allow one sodium ion and one water molecule to percolate, but too small to allow passage of hydrated potassium ions[49]. The elliptical outline of the constriction in the low-pH Δ13–PcTx1 structure is in general accord with the proposed geometry of a Na^{+}-selective pore, albeit larger, with constrictions of ~5 Å by 7 Å at Leu 440 of subunit C and of ~4 Å by 10 Å at Leu 440 of subunits A and B (Supplementary Fig. 16). The extent to which the mechanism of Na^{+}-selectivity upon proton activation, in the absence of PcTx1, hinges on Leu 440 and on the low-pH Δ13–PcTx1 conformation requires further experimentation. Nevertheless, mutation of Leu 440 to Ala or Ser (Supplementary Fig. 17) demonstrates that Leu 440 is crucial to the formation and function of the ion channel pore on the basis of the reduced activity of the mutants compared to the parent construct.

Ion-binding sites

To map ion-binding sites we soaked crystals in solutions containing Cs^+, a voltage-dependent, open channel blocker of the Δ13 construct (Supplementary Fig. 18). Inspection of anomalous difference electron density maps revealed a site (6–12σ) in each subunit, common to the high- and low-pH Δ13–PcTx1 crystal forms and located at the interface between the wrist and the extracellular end of TM1 (Fig. 5a and Supplementary Fig. 19). In both crystal forms, the backbone carbonyl oxygens of Leu 71, Tyr 72, Pro 287 and Trp 288 coordinate Cs^+ (Fig. 5b). Interestingly, the three crystallographically independent Cs^+ sites in the C2 structure vary in strength, with subunit C having the weakest signal, thus indicating that the binding of ions to these sites is influenced by the variation in TM1/wrist conformations between subunits. Because Cs^+-soaking experiments with crystals of the desensitized state failed to reveal ion binding to this site, and TM1 (subunit C) in the low-pH Δ13–PcTx1 state harbours the weakest Cs^+ site, we suggest the occupancy of the site is state-dependent, with ion binding favoured when the pore is in an open state, augmenting interactions between Tyr 72 and Trp 288.

We identified two Cs^+ sites in the electrostatically negative mouth of the pore, near Asp 433 in the *C*2 structure (Fig. 5c, d), a residue implicated in stabilization of the open state yet not crucial to ion selectivity[50]. Ions at these sites are approximately 5 Å from the closest protein residues, consistent with the notion that they are low-affinity, transiently occupied cation sites bound by water-mediated contacts. We did not observe anomalous difference density features deeper into the pore, perhaps because the hydrophobic pore is devoid of favourable binding sites and the structural analysis was carried out in the absence of a membrane potential.

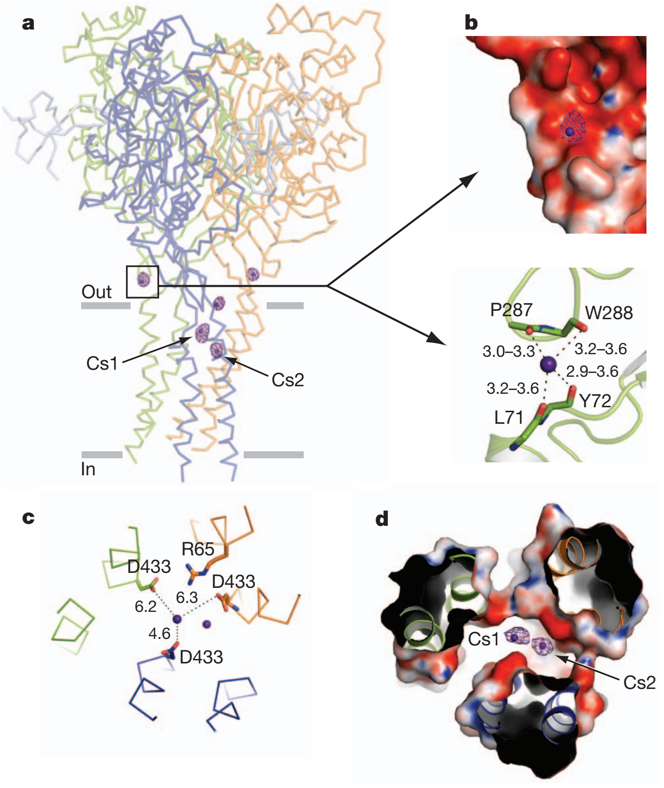


Figure 5 | Cs^+ binding sites. **a**, Anomalous difference electron density map contoured at 3.5σ. The low-pH complex is in ribbon representation. Caesium sites in the outer pore region are labelled Cs1 and Cs2. **b**, Close-up view of the Cs^+ site in the wrist region. Top, electrostatic potential contoured from −20 kT (red) to +20 kT (blue). Bottom, Cs^+ coordination by backbone carbonyl oxygens. The main chain is drawn as sticks. **c**, **d**, Close-up view of Cs^+ sites near Asp 433 shown in ribbon and stick representation (**c**) and solvent-accessible surface coloured by electrostatic potential (**d**) viewed from the extracellular side.

Mechanism

We used PcTx1 to stabilize open states of the Δ13 construct at high and low pH. Comparison of the open state and desensitized state structures defines the upper palm and knuckle as a structural scaffold and the lower palm as a conformationally flexible, proton-sensitive domain at the core of ion channel gating (Supplementary Fig. 7). The finger and thumb domains flank the palm domain, harbour binding sites for protons and PcTx1, and modulate movements of the lower palm domain by alterations in intersubunit contacts. In the open conformations, the subunit interface between the thumb and the palm domain separates while the extracellular and transmembrane domain interface forming the extracellular vestibule expands. The presence of a modulator, such as PcTx1, precludes 'collapse' of the thumb–palm subunit interface to a desensitized state conformation. The motions of the lower palm domain converge at the wrist region, inducing radial and rotational movements of the transmembrane domains, gating the ion channel (Fig. 6 and Supplementary Fig. 20).

Conclusion

Our functional and structural studies of the chicken ASIC1a complex with PcTx1 at two proton concentrations provide new insights into how the movements of the multiple domains of ASICs are coupled to ligand- and pH-dependent gating. Channel opening is correlated to the expansion of the extracellular vestibule and to rearrangements at subunit interfaces, movements coupled to the ion channel pore by the direct connections of extracellular β-strands to the transmembrane α-helices and also by the non-covalent contacts between the thumb and wrist region, a previously unrecognized site of cation binding in the open state. The non-selective and Na^+-selective states of the ion channel pore illustrate how transmembrane helices can rearrange to form a large, non-selective pore at high pH and a small, asymmetric and Na^+-selective channel at low pH. Together, these studies illuminate mechanisms of gating and selectivity in ASIC/ENaC/DEG channels.

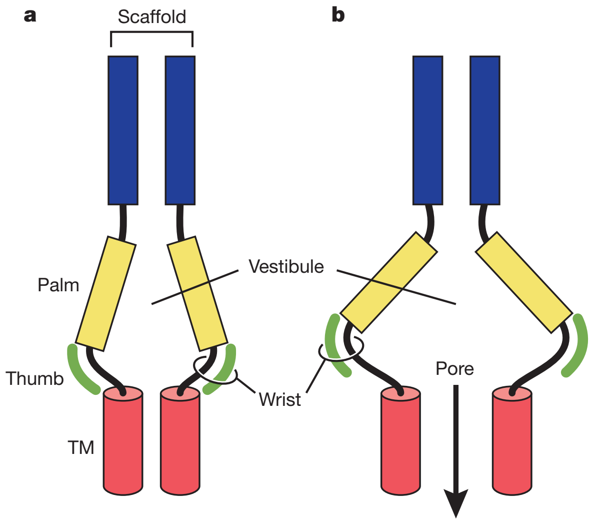


Figure 6 | Schematic representation of gating. **a**, The extracellular vestibule in the closed, desensitized state structure adopts a contracted conformation. **b**, In the open pore conformation the vestibule occupies an expanded conformation, stabilized by the thumb domain. The wrist region, in turn, couples the conformational changes of the extracellular domain to the transmembrane domains (red cylinder) of the ion channel pore.

METHODS SUMMARY

The Δ13 construct was expressed in insect or mammalian cells using baculovirus-mediated expression and purified by metal ion affinity and size-exclusion chromatography. The high- and low-pH crystal forms were obtained in conditions containing 20 mM Tris (pH 7.25) and 14–18% PEG 550 MME, or 100 mM sodium acetate (pH 5.5) and 9–12% PEG 2000 MME, respectively. Data processing, model building and refinement were performed using the HKL2000, COOT and PHENIX computer programs. The structures were solved by molecular replacement and subjected to crystallographic refinement. Whole-cell recordings were performed using CHO-K1 cells transfected with plasmid DNA encoding the Δ13 construct.

Full Methods and any associated references are available in the online version of the paper.

Received 29 April; accepted 10 July 2012.

Published online 29 July 2012.

1. Gründer, S. & Chen, X. Structure, function, and pharmacology of acid-sensing ion channels (ASICs): focus on ASIC1a. *Int. J. Physiol. Pathophysiol. Pharmacol.* **2,** 73–94 (2010).
2. Kellenberger, S. & Schild, L. Epithelial sodium channel/degenerin family of ion channels: a variety of functions for a shared structure. *Physiol. Rev.* **82,** 735–767 (2002).
3. Kashlan, O. B. & Kleyman, T. R. ENaC structure and function in the wake of a resolved structure of a family member. *Am. J. Physiol. Renal Physiol.* **301,** F684–F696 (2011).
4. Krishtal, O. A. & Pidoplichko, V. I. A receptor for protons in the nerve cell membrane. *Neuroscience* **5,** 2325–2327 (1980).
5. Waldmann, R. Proton-gated cation channels–neuronal acid sensors in the central and peripheral nervous system. *Adv. Exp. Med. Biol.* **502,** 293–304 (2001).
6. Alvarez de la Rosa, D. *et al.* Distribution, subcellular localization and ontogeny of ASIC1 in the mammalian central nervous system. *J. Physiol. (Lond.)* **546,** 77–87 (2003).
7. Wemmie, J. A. *et al.* Acid-sensing ion channel 1 is localized in brain regions with high synaptic density and contributes to fear conditioning. *J. Neurosci.* **23,** 5496–5502 (2003).
8. Deval, E. *et al.* Acid-sensing ion channels (ASICs): pharmacology and implication in pain. *Pharmacol. Ther.* **128,** 549–558 (2010).
9. Bohlen, C. J. *et al.* A heteromeric Texas coral snake toxin targets acid-sensing ion channels to produce pain. *Nature* **479,** 410–414 (2011).
10. Coryell, M. W. *et al.* Restoring acid-sensing ion channel-1a in the amygdala of knock-out mice rescues fear memory but not unconditioned fear responses. *J. Neurosci.* **28,** 13738–13741 (2008).
11. Krishtal, O. The ASICs: signaling molecules? Modulators? *Trends Neurosci.* **26,** 477–483 (2003).
12. Xiong, Z. G. *et al.* Neuroprotection in ischemia: blocking calcium-permeable acid-sensing ion channels. *Cell* **118,** 687–698 (2004).
13. Yermolaieva, O., Leonard, A. S., Schnizler, M. K., Abboud, F. M. & Welsh, M. J. Extracellular acidosis increases neuronal cell calcium by activating acid-sensing ion channel 1a. *Proc. Natl Acad. Sci. USA* **101,** 6752–6757 (2004).
14. Schild, L. The epithelial sodium channel and the control of sodium balance. *Biochim. Biophys. Acta* **1802,** 1159–1165 (2010).
15. Snyder, P. M. *et al.* Mechanism by which Liddle's syndrome mutations increase activity of a human epithelial Na^+ channel. *Cell* **83,** 969–978 (1995).
16. Chang, S. S. *et al.* Mutations in subunits of the epithelial sodium channel cause salt wasting with hyperkalaemic acidosis, pseudohypoaldosteronism type 1. *Nature Genet.* **12,** 248–253 (1996).
17. Jasti, J., Furukawa, H., Gonzales, E. & Gouaux, E. Structure of acid-sensing ion channel 1 at 1.9 Å resolution and low pH. *Nature* **449,** 316–323 (2007).
18. Canessa, C. M., Horisberger, J.-D. & Rossier, B. C. Epithelial sodium channel related to proteins involved in neurodegeneration. *Nature* **361,** 467–470 (1993).
19. Palmer, L. G. Ion selectivity of the apical membrane Na channel in the toad urinary bladder. *J. Membr. Biol.* **67,** 91–98 (1982).
20. Lingueglia, E. *et al.* A modulatory subunit of acid sensing ion channels in brain and dorsal root ganglion cells. *J. Biol. Chem.* **272,** 29778–29783 (1997).
21. de Weille, J. R., Bassilana, F., Lazdunski, M. & Waldmann, R. Identification, functional expression and chromosomal localization of a sustained human proton-gated cation channel. *FEBS Lett.* **433,** 257–260 (1998).
22. Yagi, J., Wenk, H. N., Naves, L. A. & McCleskey, E. W. Sustained currents through ASIC3 ion channels at the modest pH changes that occur during myocardial ischemia. *Circ. Res.* **99,** 501–509 (2006).
23. Yu, Y. *et al.* A nonproton ligand sensor in the acid-sensing ion channel. *Neuron* **68,** 61–72 (2010).
24. Li, W. G., Yu, Y., Huang, C., Cao, H. & Xu, T. L. Nonproton ligand sensing domain is required for paradoxical stimulation of acid-sensing ion channel 3 (ASIC3) channels by amiloride. *J. Biol. Chem.* **286,** 42635–42646 (2011).
25. Springauf, A., Bresenitz, P. & Gründer, S. The interaction between two extracellular linker regions controls sustained opening of acid-sensing ion channel 1. *J. Biol. Chem.* **286,** 24374–24384 (2011).
26. Khakh, B. S. & Lester, H. A. Dynamic selectivity filters in ion channels. *Neuron* **23,** 653–658 (1999).
27. Waldmann, R., Champigny, G., Bassilana, F., Heurteaux, C. & Lazdunski, M. A proton-gated cation channel involved in acid-sensing. *Nature* **386,** 173–177 (1997).
28. Escoubas, P. *et al.* Isolation of a tarantula toxin specific for a class of proton-gated Na^+ channels. *J. Biol. Chem.* **275,** 25116–25121 (2000).
29. Escoubas, P., Bernard, C., Lambeau, G., Lazdunski, M. & Darbon, H. Recombinant production and solution structure of PcTx1, the specific peptide inhibitor of ASIC1a proton-gated cation channels. *Protein Sci.* **12,** 1332–1343 (2003).
30. Chen, X., Kalbacher, H. & Gründer, S. The tarantula toxin psalmotoxin 1 inhibits acid-sensing ion channel (ASIC) 1a by increasing its apparent H^+ affinity. *J. Gen. Physiol.* **126,** 71–79 (2005).
31. Chen, X., Kalbacher, H. & Gründer, S. Interaction of acid-sensing ion channel (ASIC) 1 with the tarantula toxin psalmotoxin 1 is state dependent. *J. Gen. Physiol.* **127,** 267–276 (2006).
32. Samways, D. S., Harkins, A. B. & Egan, T. M. Native and recombinant ASIC1a receptors conduct negligible Ca^{2+} entry. *Cell Calcium* **45,** 319–325 (2009).
33. Mazzuca, M. *et al.* A tarantula peptide against pain via ASIC1a channels and opioid mechanisms. *Nature Neurosci.* **10,** 943–945 (2007).
34. Salinas, M. *et al.* The receptor site of the spider toxin PcTx1 on the proton-gated cation channel ASIC1a. *J. Physiol. (Lond.)* **570,** 339–354 (2006).
35. Saez, N. J. *et al.* A dynamic pharmacophore drives the interaction between psalmotoxin-1 and the putative drug target acid-sensing ion channel 1a. *Mol. Pharmacol.* **80,** 796–808 (2011).
36. Pietra, F. Docking and MD simulations of the interaction of the tarantula peptide psalmotoxin-1 with ASIC1a channels using a homology model. *J. Chem. Inf. Model.* **49,** 972–977 (2009).
37. Sherwood, T. *et al.* Identification of protein domains that control proton and calcium sensitivity of ASIC1a. *J. Biol. Chem.* **284,** 27899–27907 (2009).
38. Dawson, R. J. P. *et al.* Structure of the acid-sensin ion channel 1 in complex with the gating modifier Psalmotoxin 1. *Nature Commun.* **3,** 936 (2012).
39. Hattori, M. & Gouaux, E. Molecular mechanism of ATP binding and ion channel activation in P2X receptors. *Nature* **485,** 207–212 (2012).
40. Gonzales, E. B., Kawate, T. & Gouaux, E. Pore architecture and ion sites in acid-sensing ion channels and P2X receptors. *Nature* **460,** 599–604 (2009).
41. Cushman, K. A., Marsh-Haffner, J., Adelman, J. & McCleskey, E. W. A conformational change in the extracellular domain that accompanies desensitization of acid-sensing ion channel (ASIC) 3. *J. Gen. Physiol.* **129,** 345–350 (2007).
42. Li, T., Yang, Y. & Canessa, C. M. Asn^{415} in the β11-β12 linker decreases proton-dependent desensitization of ASIC1. *J. Biol. Chem.* **285,** 31285–31291 (2010).
43. Li, T., Yang, Y. & Canessa, C. M. Leu^{85} in the β1-β2 linker of ASIC1 slows activation and decreases the apparent proton affinity by stabilizing a closed conformation. *J. Biol. Chem.* **285,** 22706–22712 (2010).
44. Li, T., Yang, Y. & Canessa, C. M. Two residues in the extracellular domain convert a nonfunctional ASIC1 into a proton-activated channel. *Am. J. Physiol. Cell Physiol.* **299,** C66–C73 (2010).
45. Li, J., Sheng, S., Perry, C. J. & Kleyman, T. R. Asymmetric organization of the pore region of the epithelial sodium channel. *J. Biol. Chem.* **278,** 13867–13874 (2003).
46. Poët, M. *et al.* Exploration of the pore structure of a peptide-gated Na^+ channel. *EMBO J.* **20,** 5595–5602 (2001).
47. Li, T., Yang, Y. & Canessa, C. M. Outlines of the pore in open and closed conformations describe the gating mechanism of ASIC1. *Nat Commun.* **2,** 399 (2011).
48. Waldmann, R., Champigny, G., Bassilana, F., Voilley, N. & Lazdunski, M. Molecular cloning and functional expression of a novel amiloride-sensitive Na^+ channel. *J. Biol. Chem.* **270,** 27411–27414 (1995).
49. Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, 2001).
50. Li, T., Yang, Y. & Canessa, C. M. Asp^{433} in the closing gate of ASIC1 determines stability of the open state without changing properties of the selectivity filter or Ca^{2+} block. *J. Gen. Physiol.* **137,** 289–297 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We appreciate assistance in the initial characterization of the action of PcTx1 on chicken ASIC1a by D. Samways and T. Egan, mass spectrometry analysis by D. King, together with comments from C. Jahr and R. MacKinnon. We are grateful to K. C. Garcia, C. Lee and A. Goehring for assistance with protein expression in mammalian cells. We thank L. Vaskalis for assistance with figures, H. Owen for help with manuscript preparation, H. Krishnamurthy for assistance in initial data collection, and Gouaux lab members for helpful discussion. This work was supported by an individual National Research Service Award from the National Institute of Neurological Disorders and Stroke (I.B.) and by the NIH (E.G.). E.G. is an investigator with the Howard Hughes Medical Institute.

Author Contributions I.B. and E.G. designed the project. I.B. performed protein purification, crystallography and electrophysiology. I.B. and E.G. wrote the manuscript.

Author Information Structures for low-pH and high-pH Δ13–PcTx1 complexes have been deposited in PDB under accession numbers 4FZ0 and 4FZ1, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.G. (gouauxe@ohsu.edu).

METHODS

Expression and purification. The construct Δ13 was derived from the chicken *ASIC1* gene and was expressed as an N-terminal fusion with octa-histidine-tagged enhanced green fluorescent protein (EGFP) using baculovirus expression systems in insect cells and mammalian cells[51] with two thrombin sites and one thrombin site, respectively, encoded upstream of Δ13, generating a final protein sequence containing residues Gly 14 to Arg 463, verified by N-terminal amino acid sequencing. Protein expressed in insect cells was purified as described[40], whereas the mammalian cells expressing Δ13 protein were collected by centrifugation (1,000*g*) and sonicated in the presence of 150 mM NaCl, 20 mM Tris (pH 8.0) and protease inhibitors, then subsequently solubilized in 40 mM *n*-dodecyl β-D-maltoside (DDM) for 1 h at 4 °C. The solubilized material was clarified by centrifugation (19,000*g*) for 1 h at 4 °C and the supernatant was incubated with TALON resin for 1.5 h at 4 °C. Bound protein was then eluted with 150 mM NaCl, 20 mM Tris (pH 8.0), 250 mM imidazole and 1 mM DDM. Cleavage of the histidine-tagged GFP was achieved by thrombin. The resulting Δ13 protein was further purified by size-exclusion chromatography using a mobile phase containing 150 mM NaCl, 20 mM Tris (pH 7.4), 1 mM DDM, 1 mM DTT and 1 mM EDTA. Peak fractions were collected and concentrated to 1.60 mg ml^{-1} based on A_{280} measurement. Synthetic psalmotoxin (Peptides International) was immediately added in a 6:1 molar ratio of toxin to channel before crystallization.

Crystallization. Crystals of the high-pH form of the Δ13–PcTx1 complex were grown with a protein solution supplemented with 100 μM 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine (DMPC) by way of vapour diffusion using a reservoir solution composed of 20 mM Tris (pH 7.25), 14–18% PEG 550 MME and drops composed of a ratio of 1.5:1 and 2:1, reservoir to protein, respectively, at 4 °C. For cryoprotection, crystals were soaked in reservoir solution supplemented with increasing concentrations of glycerol (15% v/v final concentration), 1 mM DDM and 1 mM DMPC. For Cs^+ anomalous diffraction experiments, crystals were soaked in reservoir solutions supplemented with 300 mM CsCl, 500 μM DMPC and 1 mM DDM.

Crystals of the low-pH form of the Δ13–PcTx1 complex were obtained by vapour diffusion in hanging drop configuration using a 1:2 reservoir to protein ratio at 4 °C. The reservoir solution typically contained 100 mM sodium acetate (pH 5.5) and 9–12% PEG 2000 MME. Cryoprotection was accomplished using reservoir solution supplemented with increasing concentrations of glycerol (20% v/v final concentration). For Cs^+ anomalous diffraction experiments, a modified reservoir solution was employed in which the sodium acetate was replaced by caesium acetate, and the solution was supplemented with 500 mM CsCl and 1 mM DDM.

Structure determination. X-ray diffraction data sets from crystals grown at pH 7.25 and belonging to the *R*3 space group were collected at the Advanced Light Source (beamline 5.0.2) and diffraction was measured to ~3.35 Å resolution. The best X-ray diffraction data sets from the *C*2 form, pH 5.5 crystals, were collected at the Advanced Photon Source (beamline 24ID-C) and scaled to ~2.80 Å resolution. Diffraction data for crystals soaked in caesium-containing solutions were measured using low energy X-rays (9,000 eV) at the Advanced Light Source. Diffraction data were indexed, integrated and scaled using the HKL2000 software[52]. The structures were solved by molecular replacement using the program PHASER[53], with the extracellular domain coordinates of the ΔASIC1 structure[17] (PDB code 2QTS) as a search probe. Resulting maps after molecular replacement showed strong density of psalmotoxin, and thus the toxin was fitted into the density at the subunit interface using the solution structure[29]. Iterative model building and refinement were performed using the COOT[54] and PHENIX[55] package. Crystals in the *R*3 space group contained one subunit and one toxin in the asymmetric unit; consequently, the channel is built using the subunits and toxins related by crystallographic symmetry. Alternatively, crystals in the *C*2 space group contained one trimer and three toxins in the asymmetric unit. Manual building of the transmembrane domains was guided by electron density maps calculated using the crystallographically refined coordinates of the extracellular domain–toxin complex. 'Omit' electron density maps were calculated to validate residue registration, particularly with the transmembrane domains. Subsequent iterative cycles of model building and refinement were performed after building the transmembrane regions. Non-crystallographic symmetry (NCS) restraints were implemented in the extracellular domain and toxin for the low-pH structure containing regions defined automatically by the PHENIX package to improve maps. These regions consist of residues 72–135, 138–153, 155-291, 303–331, 333–360, 362–386 and 388–427 in the extracellular domain and residues 2–38 in the toxin. As for the high-pH structure, TLS parameters were refined with 7 TLS groups defined by PHENIX, which comprised of 5 groups in the ASIC subunit and 2 groups in the toxin. Structure validation was performed using MolProbity[56]. The low-pH final model contains channel subunits with residues 50–454, 45–450 and 42–454 of chains A, B and C, respectively, and toxins with residues 2–38, 2–37 and 1–38 of chains M, N and O, respectively. The high-pH model contains residues 41–450 of the ASIC subunit, and 2-13 and 16-38 of the toxin. A region in the thumb domain of the high-pH structure lacked electron density consisting of residues 298–301 and was omitted in the final structure. Pore surface and dimension were determined using the software HOLE[57], whereas the rotation axis shown in Fig. 3a was analysed using Dyndom[58].

Electrophysiology. Whole-cell recordings were carried out with CHO-K1 cells 24–48 h after transfection with plasmid DNA encoding the Δ13 construct and GFP expressed from an internal ribosome entry site. Pipettes were pulled and polished to 2–3 MΩ resistance and filled with internal solution containing (in mM): 150 KCl, 2 $MgCl_2$, 5 EGTA and 10 HEPES (pH 7.35). External solution contained (in mM): 150 NaCl, 2 $MgCl_2$, 2 $CaCl_2$, 8 Tris and 4 MES. For ion selectivity experiments, NaCl was substituted with equimolar concentrations of LiCl, KCl, CsCl or NMDG in the external solution. Proton concentration–response profiles were recorded and normalized to maximal current generated by application of external solution buffered at pH 5.8, and the PcTx1 concentration–response curve was generated by normalizing to maximal current generated by application of 1 μM of toxin. All concentration–response curves were fitted to the Hill equation.

51. Dukkipati, A., Park, H. H., Waghray, D., Fischer, S. & Garcia, K. C. BacMam system for high-level expression of recombinant soluble and membrane glycoproteins for structural studies. *Protein Expr. Purif.* **62,** 160–170 (2008).
52. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
53. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D* **63,** 32–41 (2007).
54. Emsley, P. & Cowtan, K. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
55. Adams, P. D. *et al. PHENIX*: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
56. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** W375–W383 (2007).
57. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Samsom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14,** 354–360 (1996).
58. Hayward, S. & Lee, R. A. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J. Mol. Graph. Model.* **21,** 181–183 (2002).

doi:10.1038/nature11446

A magnified young galaxy from about 500 million years after the Big Bang

Wei Zheng[1], Marc Postman[2], Adi Zitrin[3], John Moustakas[4], Xinwen Shu[5], Stephanie Jouvel[6,7], Ole Høst[6], Alberto Molino[8], Larry Bradley[2], Dan Coe[2], Leonidas A. Moustakas[9], Mauricio Carrasco[10], Holland Ford[1], Narciso Benítez[8], Tod R. Lauer[11], Stella Seitz[12], Rychard Bouwens[13], Anton Koekemoer[2], Elinor Medezinski[1], Matthias Bartelmann[3], Tom Broadhurst[14,15], Megan Donahue[16], Claudio Grillo[17], Leopoldo Infante[10], Saurabh W. Jha[18], Daniel D. Kelson[19], Ofer Lahav[6], Doron Lemze[1], Peter Melchior[20], Massimo Meneghetti[21], Julian Merten[9], Mario Nonino[22], Sara Ogaz[2], Piero Rosati[23], Keiichi Umetsu[24] & Arjen van der Wel[25]

Re-ionization of the intergalactic medium occurred in the early Universe at redshift $z \approx 6$–11, following the formation of the first generation of stars[1]. Those young galaxies (where the bulk of stars formed) at a cosmic age of less than about 500 million years ($z \lesssim 10$) remain largely unexplored because they are at or beyond the sensitivity limits of existing large telescopes. Understanding the properties of these galaxies is critical to identifying the source of the radiation that re-ionized the intergalactic medium. Gravitational lensing by galaxy clusters allows the detection of high-redshift galaxies fainter than what otherwise could be found in the deepest images of the sky[2]. Here we report multiband observations of the cluster MACS J1149+2223 that have revealed (with high probability) a gravitationally magnified galaxy from the early Universe, at a redshift of $z = 9.6 \pm 0.2$ (that is, a cosmic age of 490 ± 15 million years, or 3.6 per cent of the age of the Universe). We estimate that it formed less than 200 million years after the Big Bang (at the 95 per cent confidence level), implying a formation redshift of $\lesssim 14$. Given the small sky area that our observations cover, faint galaxies seem to be abundant at such a young cosmic age, suggesting that they may be the dominant source for the early re-ionization of the intergalactic medium.

Galaxy clusters are the largest reservoirs of gravitationally bound dark matter, and their huge masses bend light and form 'cosmic lenses'. They can significantly increase the brightnesses and sizes of galaxies far beyond them, thereby revealing morphological details that are otherwise impossible to detect[3–8] and allowing the spectroscopic study of the physical conditions in intrinsically faint galaxies. Most galaxies at $z \approx 10$ are expected to be fainter than an AB magnitude (the system used throughout the text) of ~29 mag (refs 9–11), which is below the imaging detection limits of the deepest fields observed by NASA's Hubble Space Telescope (HST), and largely beyond the spectroscopic capability of even the next generation of large telescopes. A gain in sensitivity through gravitational lensing is particularly valuable for the infrared data collected by NASA's Spitzer Space Telescope because the telescope's low spatial resolution blends faint sources and hampers extremely deep observations.

By combining HST and Spitzer data we are able to estimate the age of such distant objects on the basis of the ratio of their rest-frame ultraviolet to optical fluxes. The age and distance estimates rely, in large part, on measuring the observed wavelengths and relative amplitudes of prominent hydrogen absorption features in the spectra of faint galaxies. At $z > 7$, the hydrogen Lyman α break, at a wavelength of $\sim 0.12(1 + z)$ μm, is redshifted out of the optical bands, and the hydrogen Balmer break, at $\sim 0.38(1 + z)$ μm, is redshifted into the range of the Infrared Array Camera (IRAC) on board Spitzer.

We have discovered a gravitationally lensed source whose most likely redshift is $z \approx 9.6$. The source, hereafter called MACS 1149-JD, is selected from a near-infrared detection image at a significance of 22σ. MACS 1149-JD has a unique flux distribution characterized by no detection at wavelengths shorter than 1.2 μm, firm detections in the two reddest HST bands, and weak detections in two bands of the HST Wide-Field Camera 3/Infrared Channel and in one IRAC channel (Fig. 1). The object's coordinates (J2000) are RA = 11 h 49 min 33.584 s, dec. = +22° 24′ 45.78″ (Fig. 2).

The Cluster Lensing And Supernova survey with Hubble (CLASH)[12] is an HST Multi-Cycle Treasury programme that is

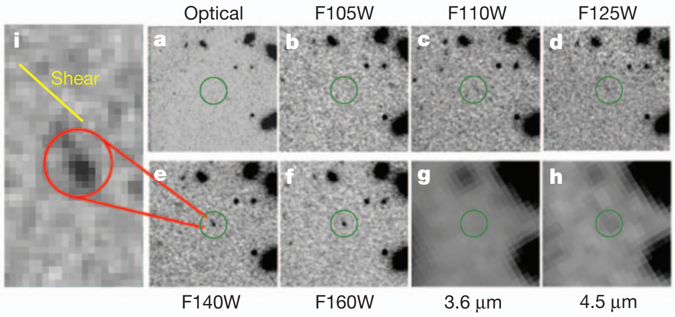


Figure 1 | Multiband images of the $z = 9.6$ galaxy candidate MACS 1149-JD. a–h, The optical image (**a**) shows the sum of all data from the HST Advanced Camera for Surveys. The source, located at the centre of each image (green circle), is clearly detected in the F140W (central wavelength, 1.39 μm; **e**) and F160W (1.53 μm; **f**) bands and weakly detected in the F110W (1.15 μm; **c**), F125W (1.25 μm; **d**) and 4.5-μm (**h**) bands. The F105W band (1.05 μm; **b**) extends to ~1.2 μm, and the lack of detection in that band confirms that there is no source flux below that wavelength. The source is not detected in the 3.6-μm band (**g**). Each of these images is 10 arcsec on each side. North is up and east is to the left. **i**, An enlarged view of the F140W image (**e**) shows the elongation of the source, which is extended along a position angle of ~47°. The yellow line marks the direction of shear predicted by the lensing model, and the red circle marks the aperture used for the source photometry (ten pixels in diameter).

[1] Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA. [2] Space Telescope Science Institute, Baltimore, Maryland 21218, USA. [3] Institut für Theoretische Astrophysik, Universität Heidelberg, 69120 Heidelberg, Germany. [4] Department of Physics and Astronomy, Siena College, Loudonville, New York 12211, USA. [5] Department of Astronomy, University of Science and Technology of China, Hefei, Anhui 230026, China. [6] Department of Physics and Astronomy, University College London, London WC1E 6BT, UK. [7] Institute de Ciencies de l'Espai, 08193 Bellaterra, Spain. [8] Instituto de Astrofísica de Andalucía, E-18008 Granada, Spain. [9] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. [10] Departamento de Astronomía y Astrofísica, Pontificia Universidad Católica de Chile, Santiago 22, Chile. [11] National Optical Astronomical Observatory, Tucson, Arizona 85726, USA. [12] Universitäts-Sternwarte München, D-81679 München, Germany. [13] Leiden Observatory, Leiden University, 2300 RA Leiden, The Netherlands. [14] Department of Theoretical Physics, University of Basque Country, 48080 Bilbao, Spain. [15] Ikerbasque, Basque Foundation for Science, 48011 Bilbao, Spain. [16] Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA. [17] Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark. [18] Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA. [19] The Observatories of the Carnegie Institution for Science, Pasadena, California 91101, USA. [20] Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA. [21] INAF-Osservatorio Astronomico di Bologna, I-40127 Bologna, Italy. [22] INAF-Osservatorio Astronomico di Trieste, 40131 Trieste, Italy. [23] European Southern Observatory, D-85748 Garching, Germany. [24] Academia Sinica, Institute of Astronomy and Astrophysics, Taipei 10617, Taiwan. [25] Max-Planck Institut für Astronomie, D-69117 Heidelberg, Germany.

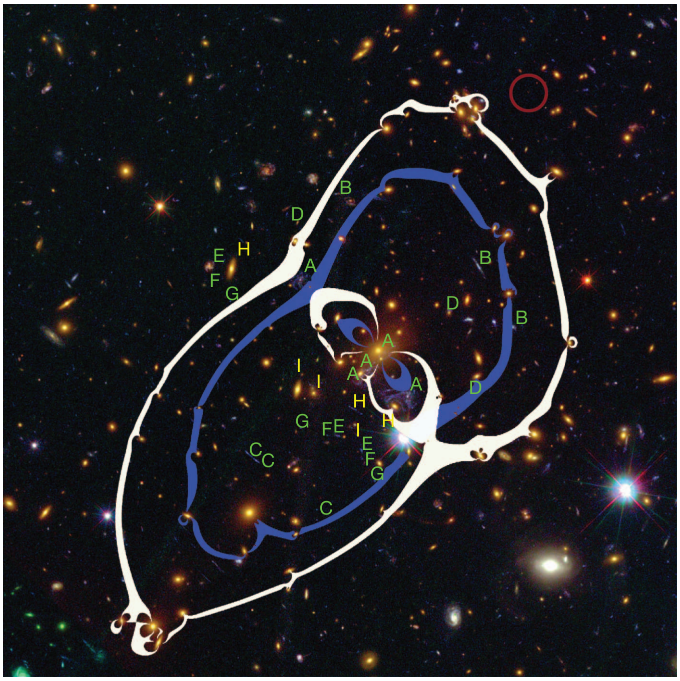


Figure 2 | Composite colour image of MACS J1149.6+2223 made from multiband data. North is up and east is to the left. The field of view is 2.2 arcmin on each side. The $z = 9.6$ critical curve for the best-fit lensing model is overlaid in white, and that for $z = 3$ is shown in blue. Green letters A–G mark the multiple images of seven sources that were used in the strong-lensing model. Yellow letters H and I mark the two systems that were not used in the final fitting. The location of MACS 1149-JD, at RA = 11 h 49 min 33.584 s, dec. = +22° 24′ 45.78′′ (J2000), is marked with a red circle.

acquiring images in 16 broad bands between 0.2 and 1.7 μm for 25 clusters. MACS J1149.6+2223 is a massive cluster ($\sim 2.5 \times 10^{15}$ solar masses ($M_\odot$)) at redshift $z = 0.544$, selected from a sample of X-ray-luminous clusters[13]. The mass models for this cluster[14,15] suggest a relatively flat mass distribution profile and a large area of high magnification, making it one of the most powerful cosmic lenses known.

The spectral energy distribution (SED) features of galaxies, most notably the Lyman break and the Balmer break, generate distinct colours between broad bands and enable us to derive their redshifts with reasonable accuracy. Our photometric redshift estimates are made using two different techniques: LE PHARE[16] (LPZ) and the Bayesian photometric redshift[17] (BPZ) method. LPZ photometric redshifts are based on a template-fitting procedure with a maximum-likelihood (χ^2) estimate. We use the template library of the COSMOS survey[18], including templates of three elliptical galaxies, seven spiral galaxies[19] and twelve galaxies in the library of Bruzual and Charlot[20], with starburst ages ranging from 30 million years (Myr) to 3 billion years (Gyr) to reproduce the bluest galaxies better. The LPZ solution from the marginalized posterior is $z = 9.60^{+0.20}_{-0.28}$ (68% confidence level), and the best-fit model is a starburst galaxy.

BPZ multiplies the likelihood by the prior probability of a galaxy with an apparent magnitude m_0 having a redshift z and spectral type T. We run BPZ using a new library composed of 11 SED templates originally drawn from PEGASE[21] but recalibrated using the FIREWORKS photometry and spectroscopic redshifts[22] to optimize its performance. This galaxy library includes five templates for ellipticals, two for spirals and four for starbursts. The most likely BPZ solution is a starburst galaxy at $z = 9.61^{+0.14}_{-0.13}$ (1σ).

Even though the CLASH data have more bands than other HST projects, MACS 1149-JD is detected only in the four reddest HST bands. The high confidence of our high-redshift solution is due to the IRAC photometry at 3.6 and 4.5 μm. With HST data alone (that is, excluding Spitzer data), solutions with intermediate redshifts ($2 \lesssim z \lesssim 6$) can be found but they have low probability (Fig. 3). When Spitzer data are included, no viable solutions other than those at $z \approx 9.6$ are found, and the possibility for photometric redshifts of $z < 8.5$ is rejected at the 4σ confidence level ($<3 \times 10^{-5}$).

Using confirmed multiply lensed images, strong-lensing models[14,23] allow us to derive the mass distribution of dark matter in the cluster, which leads to an amplification map for background sources. With 23 multiply lensed images of seven sources, some of which are large enough to comprise distinctive knots used as additional constraints, we derive the best-fit model in which the critical curve (defining regions of high magnification) for $z \approx 10$ extends into the vicinity of MACS 1149-JD, resulting in a magnification of $\mu = 14.5^{+4.2}_{-1.0}$. The results are in rough agreement with a second, independent, model[24], which yields a best-fit magnification with large error bars, $26.6^{+20.8}_{-7.7}$.

Because our data cover a broad spectral range in the object's rest frame, we are able to estimate some key properties for the source using the Bayesian SED-fitting code iSEDfit[25] coupled to state-of-the-art models of synthetic stellar populations[26] and based on the Chabrier[27] initial mass function from 0.1 $M_\odot$ to 100 $M_\odot$. We consider a wide range of parameterized star formation histories and stellar metallicities and assume no dust attenuation in our fiducial modelling (but see Supplementary Information), because previous studies[28,29] found no evidence for dust in galaxies at the highest redshifts.

Figure 4 presents the results of our population synthesis modelling adopting $z = 9.6$ as the source redshift. Based on the median of the posterior probability distributions, our analysis suggests a stellar mass of $\sim 1.5 \times 10^8 (\mu/15)^{-1} M_\odot$ and a star formation rate (SFR) of $\sim 1.2(\mu/15)^{-1} M_\odot\ \mathrm{yr}^{-1}$. We note that these values would be a factor of up to eight higher if dust attenuation were not negligible (Supplementary Information). Given the uncertainties in the IRAC photometry, we are unable to measure the age of the galaxy precisely; however, we can constrain its SFR-weighted age, or the age at which most of the stars formed, to $\langle t \rangle_{\mathrm{SFR}} < 200$ Myr (95% confidence level), suggesting a likely formation redshift of $z_f < 14.2$. Given that the source is brighter at 4.5 μm than at 3.6 μm, the presence of a Balmer break is likely, suggesting that MACS 1149-JD may not be extremely young. The age $\langle t \rangle_{\mathrm{SFR}}$ is generally consistent with the estimated ages ($\gtrsim 100$ Myr) of galaxies at slightly lower redshifts, $z \approx 7$–8 (ref. 30).

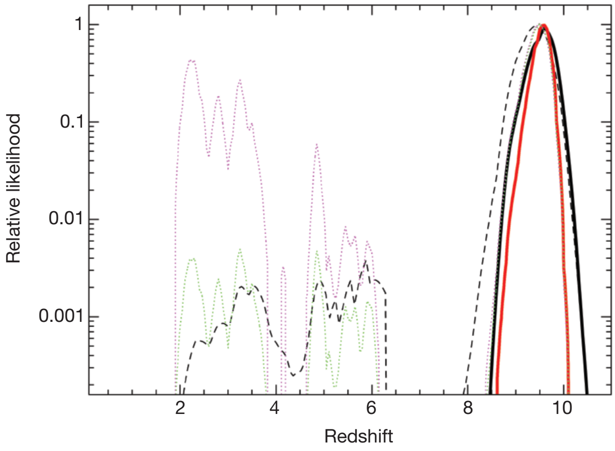


Figure 3 | Probability distributions of photometric redshift estimation. All curves are normalized to their peak probability. The solid black curve shows the LPZ result, calculated using all the HST and Spitzer data. The solid red curve shows the BPZ result calculated with and without priors, using all data. Only the high-redshift solutions are confirmed with high confidence (>4σ). The dashed black curve shows the LPZ result calculated using the HST data only. The dotted green curve shows the BPZ result calculated without priors, using the HST data only. In these two cases, intermediate-redshift solutions are present at low probability (<1%). The dotted magenta curve shows the BPZ result calculated with priors, using the HST data only. Only in such cases do intermediate-redshift solutions become significant.

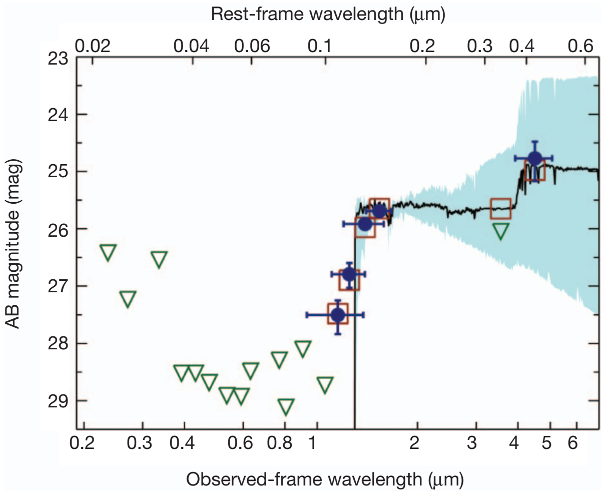


Figure 4 | Stellar population synthesis modelling results for MACS 1149-JD. The blue points mark bands in which the object is detected, and the green triangles indicate 1σ upper limits. The horizontal bars on the blue data points mark the wavelength range of the bands, and the vertical bars on the blue data points mark the 1 σ measurement errors. The errors in the F140W and F160W bands are small (<0.1 magnitude) and, hence, not visible. The black spectrum is the best-fit model, and the red squares show the photometry of this model convolved with the Wide-Field Camera 3, Advanced Camera for Surveys and IRAC filter response functions. The blue shading shows the range of 100 additional models drawn from the posterior probability distribution that are also statistically acceptable fits to the data.

On the basis of one such discovery over 12 clusters and our corresponding lensing models, we estimate that the SFR density at $z = 9$–10 is $1.8^{+4.3}_{-1.1} \times 10^{-3} M_{\odot}\,\mathrm{Mpc}^{-3}\,\mathrm{yr}^{-1}$, which is lower than the extrapolation from lower redshifts[9] but higher than the limit derived from the Hubble Ultra Deep Field[9]. Statistically, our estimate is consistent with both values, and more data are needed to reduce the uncertainties. Our current observations, coupled with knowledge of the galaxy luminosity function[1], suggest that faint galaxies at $z \approx 10$ may be the dominant source of the radiation responsible for the early re-ionization of the intergalactic medium.

Received 6 April; accepted 26 July 2012.

1. Robertson, B. E., Ellis, R. S., Dunlop, J. S., McLure, R. J. & Stark, D. P. Early star-forming galaxies and the reionization of the Universe. *Nature* **468,** 49–55 (2010).
2. Kneib, J.-P. & Natarajan, P. Cluster lenses. *Astron. Astrophys. Rev.* **19,** 47–146 (2011).
3. Kneib, J.-P., Ellis, R. S., Santos, M. R. & Richard, J. A probable $z \sim 7$ galaxy strongly lensed by the rich cluster A2218: exploring the dark ages. *Astrophys. J.* **607,** 697–703 (2004).
4. Bradley, L. D. *et al.* Discovery of a very bright strongly lensed galaxy candidate at $z \sim 7.6$. *Astrophys. J.* **678,** 647–654 (2008).
5. Zheng, W. *et al.* Bright strongly lensed galaxies at redshift $z \sim 6$–7 behind the clusters Abell 1703 and CL0024+16. *Astrophys. J.* **697,** 1907–1917 (2009).
6. Richard, J. *et al.* Discovery of a possibly old galaxy at $z = 6.027$, multiply imaged by the massive cluster Abell 383. *Mon. Not. R. Astron. Soc.* **414,** L31–L35 (2011).
7. Bradley, L. D. *et al.* Through the looking glass: bright, highly magnified galaxy candidates at $z \sim 7$ behind A1703. *Astrophys. J.* **747,** 3 (2012).
8. Zitrin, A. *et al.* CLASH: discovery of a bright $z \approx 6.2$ dwarf galaxy quadruply lensed by MACS J0329.6–0211. *Astrophys. J.* **747,** L9 (2012).
9. Bouwens, R. J. *et al.* A candidate redshift $z \approx 10$ galaxy and rapid changes in that population at an age of 500 Myr. *Nature* **469,** 504–507 (2011).
10. Oesch, P. A. *et al.* Expanded search for $z \sim 10$ galaxies from HUDF09, ERS, and CANDELS data: evidence for accelerated evolution at $z > 8$? *Astrophys. J.* **745,** 110 (2012).
11. Zackrisson, E. *et al.* Detecting gravitationally lensed population III galaxies with HST and JWST. Preprint at http://arxiv.org/abs/1204.0517 (2012).
12. Postman, M. *et al.* Cluster lensing and supernova survey with Hubble (CLASH): an overview. *Astrophys. J.* **199** (suppl.), 25 (2012).
13. Ebeling, H., Edge, A. C. & Henry, J. P. MACS: a quest for the most massive galaxy clusters in the universe. *Astrophys. J.* **553,** 668–676 (2001).
14. Zitrin, A. & Broadhurst, T. Discovery of the largest known lensed images formed by a critically convergent lensing cluster. *Astrophys. J.* **703,** L132–L136 (2009).
15. Smith, G. P. *et al.* Hubble space telescope observations of a spectacular new strong-lensing galaxy cluster: MACS J1149.5+2223 at $z = 0.544$. *Astrophys. J.* **707,** L163–L168 (2009).
16. Ilbert, O. *et al.* Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. *Astron. Astrophys.* **457,** 841–856 (2006).
17. Benítez, N. Bayesian photometric redshift estimation. *Astrophys. J.* **536,** 571–583 (2000).
18. Koekemoer, A. M. *et al.* The COSMOS survey: Hubble space telescope advanced camera for surveys observations and data processing. *Astrophys. J.* **172** (suppl.), 196–202 (2007).
19. Polletta, M. *et al.* Spectral energy distributions of hard X-ray selected active galactic nuclei in the XMM-Newton medium deep survey. *Astrophys. J.* **663,** 81–102 (2007).
20. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344,** 1000–1028 (2003).
21. Fioc, M. & Rocca-Volmerange, B. PEGASE: a UV to NIR spectral evolution model of galaxies. application to the calibration of bright galaxy counts. *Astron. Astrophys.* **326,** 950–962 (1997).
22. Wuyts, S. *et al.* FIREWORKS U_{38}-to-24 µm photometry of the GOODS Chandra deep field-south: multiwavelength catalog and total infrared properties of distant K_s-selected galaxies. *Astrophys. J.* **682** (suppl.), 985–1003 (2008).
23. Zitrin, A., Broadhurst, T., Barkana, R., Rephaeli, Y. & Benítez, N. Strong-lensing analysis of a complete sample of 12 MACS clusters at $z > 0.5$: mass models and Einstein radii. *Mon. Not. R. Astron. Soc.* **410,** 1939–1956 (2011).
24. Jullo, E. *et al.* A Bayesian approach to strong lensing modelling of galaxy clusters. *N. J. Phys.* **9,** 447 (2007).
25. Moustakas, J. *et al.* Evolution of the stellar mass-metallicity relation since $z = 0.75$. *Astrophys. J.* (submitted); preprint at http://arxiv.org/abs/1112.3300 (2011).
26. Conroy, C. & Gunn, J. E. The propagation of uncertainties in stellar population synthesis modeling. III. model calibration, comparison, and evaluation. *Astrophys. J.* **712,** 833–857 (2010).
27. Chabrier, G. Galactic stellar and substellar initial mass function. *Publ. Astron. Soc. Pacif.* **115,** 763–795 (2003).
28. Labbé, I. *et al.* Ultradeep infrared array camera observations of sub-L* $z \sim 7$ and $z \sim 8$ galaxies in the Hubble ultra deep field: the contribution of low-luminosity galaxies to the stellar mass density and reionization. *Astrophys. J.* **708,** L26–L31 (2010).
29. Bouwens, R. J. *et al.* Very blue UV-continuum slope β of low luminosity $z \sim 7$ galaxies from WFC3/IR: evidence for extremely low metallicities? *Astrophys. J.* **708,** L69–L73 (2010).
30. Labbé, I. *et al.* Star formation rates and stellar masses of $z = 7$–8 galaxies from IRAC observations of the WFC3/IR early release science and the HUDF fields. *Astrophys. J.* **716,** L103–L108 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements The CLASH programme (GO-12065) is based on observations made with the NASA/ESA Hubble Space Telescope. The Space Telescope Science Institute is operated by the Association of Universities for Research in Astronomy, Inc. under NASA contract NAS 5-26555. This work is also based in part on archival data obtained with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. J.M. acknowledges support from NSF grant AST-0908246. A. Z. is supported by research contract Internationale Spitzenforschung II-1 of the Baden Württemberg Stiftung.

Author Contributions W.Z. made the initial identification and wrote a draft. R.B., D.C., H.F. and L.B. verified the target selection. M.P and H.F performed comparisons with intermediate-redshift and nearby objects and edited the final version. W.Z., A.K., L.B., D.C., S.O. and E.M. processed the HST data. X.S., W.Z. and L.A.M. performed the IRAC photometry. S.J., A.M., D.C., O.H. and N.B. made the redshift estimates. M.P., T.R.L. and L.B. performed the image deconvolution. J.M. carried out the SED fitting. A.Z., M.C. and T.B. constructed the lensing models. L.A.M. and D.C. estimated the SFR density at $z \approx 10$. The above authors also contributed the text and figures that describe their analyses. P.R., L.I., P.M., M.N., R.B. and L.A.M. contributed to the observing programmes. M.B., M.D., D.D.K., O.L., K.U. and A.v.d.W. were involved in designing the project, reviewing the results and editing the manuscript. C.G., S.W.J., D.L. and P.M. edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.Z. (zheng@pha.jhu.edu).

doi:10.1038/nature11437

Pulsed electron paramagnetic resonance spectroscopy powered by a free-electron laser

S. Takahashi[1,2], L.-C. Brunel[2], D. T. Edwards[3], J. van Tol[4], G. Ramian[2], S. Han[2,5] & M. S. Sherwin[2,3]

Electron paramagnetic resonance (EPR) spectroscopy interrogates unpaired electron spins in solids and liquids to reveal local structure and dynamics; for example, EPR has elucidated parts of the structure of protein complexes that other techniques in structural biology have not been able to reveal[1–4]. EPR can also probe the interplay of light and electricity in organic solar cells[5–7] and light-emitting diodes[8], and the origin of decoherence in condensed matter, which is of fundamental importance to the development of quantum information processors[9–13]. Like nuclear magnetic resonance, EPR spectroscopy becomes more powerful at high magnetic fields and frequencies, and with excitation by coherent pulses rather than continuous waves. However, the difficulty of generating sequences of powerful pulses at frequencies above 100 gigahertz has, until now, confined high-power pulsed EPR to magnetic fields of 3.5 teslas and below. Here we demonstrate that one-kilowatt pulses from a free-electron laser can power a pulsed EPR spectrometer at 240 gigahertz (8.5 teslas), providing transformative enhancements over the alternative, a state-of-the-art ~30-milliwatt solid-state source. Our spectrometer can rotate spin-1/2 electrons through π/2 in only 6 nanoseconds (compared to 300 nanoseconds with the solid-state source). Fourier-transform EPR on nitrogen impurities in diamond demonstrates excitation and detection of EPR lines separated by about 200 megahertz. We measured decoherence times as short as 63 nanoseconds, in a frozen solution of nitroxide free-radicals at temperatures as high as 190 kelvin. Both free-electron lasers and the quasi-optical technology developed for the spectrometer are scalable to frequencies well in excess of one terahertz, opening the way to high-power pulsed EPR spectroscopy up to the highest static magnetic fields currently available.

The spectral resolution, spin polarization, sensitivity, and time resolution of pulsed EPR all increase with increasing static magnetic field and the associated Larmor precession frequency (Fig. 1a)[14]. An additional advantage of high magnetic fields has recently been demonstrated at 8.5 T (Larmor precession frequency $f_{\mathrm{Larmor}} = 240$ GHz): because of the 11.5 K Zeeman temperature ($= hf_{\mathrm{Larmor}}/k_{\mathrm{B}}$, where h is Planck's constant and k_{B} is Boltzmann's constant), the spin polarization varies from >99% at 2 K to <2% at 300 K, so that the spin decoherence time T_2 increases markedly with decreasing temperature in a large class of spin systems[9,13]. In addition to the advantages of high fields and frequencies, pulsed EPR benefits from high microwave power. In a fixed geometry, the time required to rotate an ensemble of spins by a given angle, say π/2, is proportional to the strength of the resonant microwave magnetic field B_1, and hence to the square root of the power. For many systems of interest—for example, spin-labelled proteins above the 200 K protein glass transition, where they explore their biologically relevant conformational space, or spins in semiconductors (and hence electronic devices) above 20 K—spin relaxation times are much shorter than ~1 μs (refs 15, 16), and cannot be measured using state-of-the-art solid state sources at 240 GHz which typically have powers of ~30 mW. Thus, it is imperative to have much higher microwave power, allowing π/2 spin rotations in only a few nanoseconds.

Electromagnetic sources are the bottleneck in the development of high-power pulsed EPR at high magnetic fields. Most high-power pulsed EPR spectrometers operate at 0.34 T ($f_{\mathrm{Larmor}} = 9.5$ GHz) or 1.2 T ($f_{\mathrm{Larmor}} = 34$ GHz) and are powered by vacuum electronic devices called travelling wave tube amplifiers. The highest magnetic fields at which spectrometers with few-nanosecond π/2 pulses have been constructed is 3.5 T ($f_{\mathrm{Larmor}} = 95$ GHz). The 95-GHz instruments, pioneered at Cornell University and, more recently, implemented at the University of St Andrews, UK, are powered by more exotic vacuum electronic devices called extended-interaction klystron amplifiers with peak output powers of ~1 kW (refs 17, 18). Amplifiers capable of kilowatt output powers at frequencies above 200 GHz have been envisioned[19,20] but are beyond current technology.

Here we present results from a pulsed EPR spectrometer powered by a free-electron laser (FEL). The spectrometer operates at 8.5 T (240 GHz), and can be powered either by a 30-mW, 240-GHz solid-state source or by UCSB's millimetre-wave FEL (mm-FEL)[21], with a maximum power of 1 kW.

An overview of our FEL-powered pulsed EPR system is shown in Fig. 1b (see Supplementary Information for details of the set-up and the performance). The mm-FEL[21] emits pulses a few microseconds long with peak powers of several hundred watts to several kilowatts. In 240-GHz pulsed EPR operation, the FEL output is injection-locked by a stable 240-GHz solid-state source to generate few-microsecond-long pulses with a Fourier-transform limited linewidth <1 MHz (ref. 22). By cavity-dumping the FEL, the several-hundred-watt pulses can be stepped up to 1 kW for the last 40 ns, with ~10-ns rise and fall times (Fig. 1b)[23].

In the FEL-powered spectrometer, light-activated shutters 'slice' one or two pulses directly from the >100-W beam, each with a variable duration as short as 1 ns. This is a very different mode of operation to that used in other high-power pulsed EPR spectrometers, where pulse sequences are generated at low power and then amplified. The shutters are high-purity Si wafers with thicknesses carefully polished down to 1/2 wavelength[24,25]. In our 'pulse slicer' (see Fig. 1b and Supplementary Information), 'reflection switches' are Si wafers at Brewster's angle that reflect less than 1 p.p.m. in the quiescent state. When excited with 532-nm (green) light of a Nd:YAG laser with a fluence >1 mJ cm^{-2}, the reflectance rises to >60% with a sub-nanosecond rise time. The reflectance remains high for at least ~1 μs. To make a pulse of variable duration, the 240-GHz beam is directed from a reflection switch into a 'transmission switch' consisting of back-to-back corrugated horns with a Si wafer between them, illustrated in Fig. 1b. For a single assembly, the transmittance drops from near unity in the quiescent state to less than 10^{-4} when the Si is illuminated with a pulse of green light. The duration of the pulse exiting the transmission switch is determined by the optical or electronic delay between the green pulses which activate the reflection and transmission switches. Transmittance

[1] Department of Chemistry, University of Southern California, Los Angeles, California 90089, USA. [2] Institute for Terahertz Science and Technology, University of California, Santa Barbara, California 93106, USA. [3] Department of Physics, University of California, Santa Barbara, California 93106, USA. [4] National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32310, USA. [5] Department of Chemistry and Biochemistry, University of California, Santa Barbara, California 93106, USA.

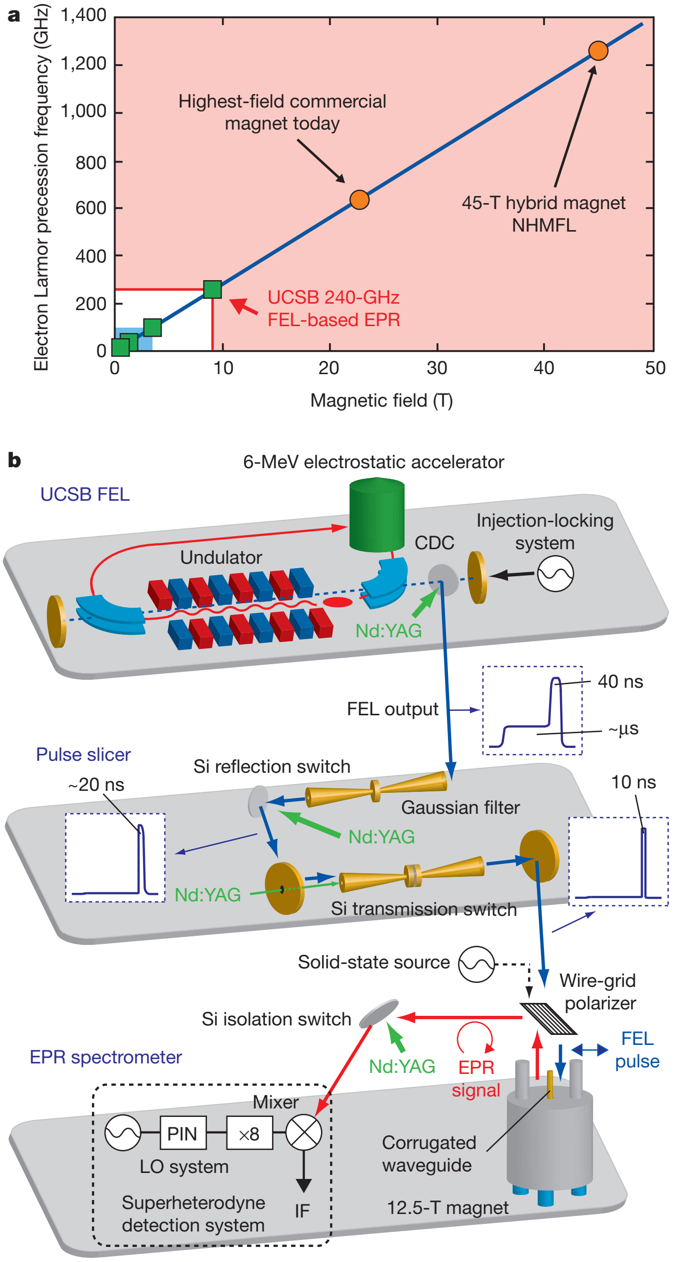


Figure 1 | Towards high-frequency, high-power pulsed EPR spectroscopy. **a**, Electron Larmor precession frequencies of existing high-power ($\sim$1 kW) EPR spectrometers and their corresponding magnetic fields (green boxes; our set-up is indicated by the red arrow, and is driven by an FEL, unlike the others shown). Existing magnet systems are shown as orange circles. The highest-field d.c. magnet is available at the National High Magnetic Field Laboratory (NHMFL). The blue line represents EPR frequencies as a function of magnetic field for $S = 1/2$ and $g = 2$ systems, where g is the electron spin g-factor. The blue shaded area shows the range of frequency and field covered by existing high-power pulsed EPR spectrometers that do not use a FEL. The horizontal red line marks the 240-GHz frequency of our EPR spectrometer, the vertical red line shows the corresponding magnetic field for systems with $S = 1/2$ and $g = 2$. The pink shaded area shows that FELs, which are tunable to frequencies well in excess of 1 THz, can enable high-power pulsed EPR spectroscopy over a field range limited only by the highest available d. c. magnetic fields. **b**, Schematic overview of the FEL-powered pulsed EPR system. The UCSB mm-FEL (see main text) emits a several-microsecond-long pulse. For control of frequency and output power, an injection-locking system and a cavity-dump coupler (CDC) have been implemented. The Nd:YAG laser-controlled pulse-slicer functions to 'slice' one or two pulses, each with variable duration as short as 1 ns, from the FEL radiation. The pulsed EPR spectrometer operates quasi-optically. For the c.w. and low-power EPR mode, the system is operated with a solid-state source at 240 GHz. The linearly polarized output of the pulse-slicer is propagated into the centre of the 12.5-T superconducting magnet in the EPR spectrometer where the nanosecond FEL pulses couple to the sample. The EPR signals are detected by a superheterodyne detection system after turning on the Si isolation switch, which protects the mixer during the pulse. See main text for details.

switches are cascaded to achieve transmittances well below 10^{-6}. A non-trivial arrangement of reflection and transmittance switches enables the generation of pairs of pulses separated by a time that can be varied from a few nanoseconds up to the several-microsecond duration of the FEL pulse (see Supplementary Information). This Si switch technology is broadband, and has been demonstrated up to frequencies of at least 30 THz (ref. 26). A pulse slicer similar to the one developed here could be used for EPR with other quasi-continuous-wave (quasi-c.w.) sources like gyrotrons[27].

The output of the pulse slicer (with a power $>$100 W) or of a 30-mW solid-state source enters a quasi-optical spectrometer very similar in design to systems that have been described previously[14,28,29]. After passing through a wire-grid polarizer, 240 GHz pulses travel down an 18-mm-diameter corrugated waveguide to a taper that reduces the beam size to either 5 or 2 mm. Solid samples are mounted directly on top of a silver-coated mirror and frozen liquid samples contained in a small Teflon bucket on the mirror. The samples are then placed at the exit of the taper in a 12.5-T superconducting magnet. No resonator is used to take advantage of large sample volume, insignificant cavity deadtime and the versatility to use various types of samples. A 240-GHz excitation pulse tips the spins to a plane perpendicular to the static magnetic field. The spins precess at a frequency near 240 GHz, emitting a sub-microwatt circularly polarized EPR signal. The mirror underneath the sample reflects both the EPR signal and the linearly polarized excitation pulse. We detect EPR signals with a superheterodyne receiver that uses a subharmonically-pumped Schottky diode as the mixer. The mixer is isolated from the excitation pulse and subsequent reflections by the induction mode isolator[28] (in which a wire grid polarizer is oriented to reflect 50% of the EPR signal but only 0.3% of the excitation pulse toward the mixer) and the isolation switch (which reflects only 1 p.p.m. before activation by a laser pulse).

In order to characterize the FEL-powered spectrometer, we used a 1:1 crystalline complex of α,γ-bisdiphenylene-β-phenylallyl (BDPA) and benzene. BDPA is a stable radical and is widely used in EPR and related techniques. The sample dimension is much smaller than the 1.25-mm wavelength of 240-GHz radiation. The exchange-narrowed 0.3-mT-wide EPR signal was measured by c.w. EPR using a solid-state source. The magnetic field was set to the centre of the EPR peak. A 10-ns pulse sliced from the FEL was applied and, after 80 ns, the isolation switch was activated. What is observed is the decaying oscillation of the 'free-induction decay' (FID) of BDPA, mixed down to an intermediate frequency (IF = 500 MHz) measurable by a transient digitizer (Fig. 2a). The intensity of the FID is then recorded by taking a fast-Fourier transform (FFT) of the FID and determining the signal intensity. In Fig. 2b, the signal intensity is plotted as a function of pulse duration varying from 0 to 30 ns in 1-ns increments. The signal undergoes Rabi oscillations[30], rising to its first maximum near 5-ns pulse duration. At the first maximum, most spins have rotated by $\pi/2$. The frequency of oscillation is $\sim$40 MHz (measurement 1 in Fig. 2b), corresponding to a $\pi/2$ pulse of $\sim$6 ns and a transverse magnetic field B_1 of $\sim$3.0 mT at the sample. The minima of the Rabi oscillations at 12 and 25 ns are not zero, indicating a significant inhomogeneity in B_1. The origin of this inhomogeneity is under investigation. The 6-ns $\pi/2$ pulse is 50 times shorter than can be achieved in this and similar spectrometers for a spin-1/2 system with a lower-power solid-state source.

The short excitation pulses enable Fourier transform EPR of single nitrogen (N) impurity spins in diamond. As shown in Fig. 3a, a N impurity exists in a diamond lattice as a substitutional impurity, and has one unpaired electron with spin $S = 1/2$. Because of the hyperfine coupling between the N electron spin and the ^{14}N nuclear spin ($I = 1$), the energy levels of the N electron spin are split into six levels, as

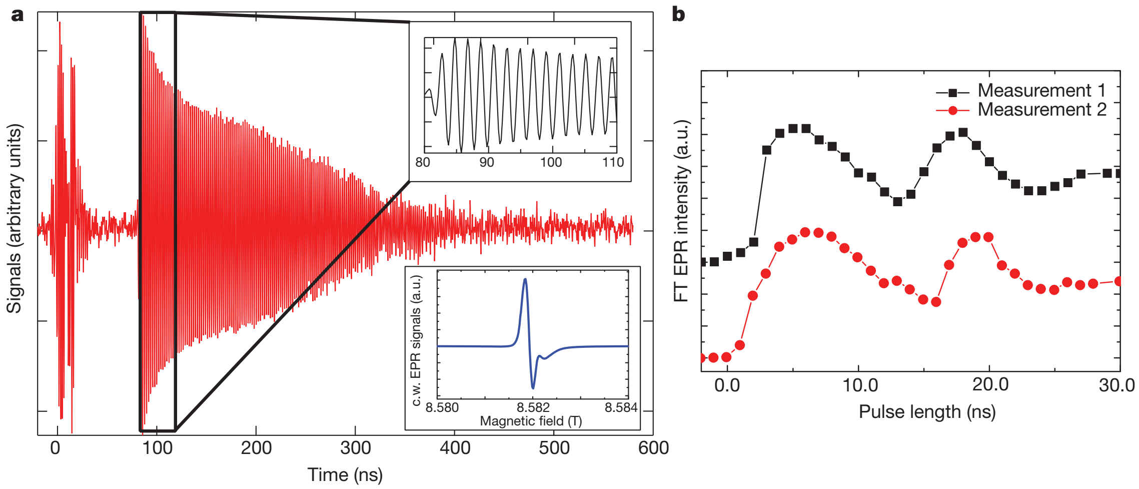


Figure 2 | Rabi oscillation measurements with BDPA. a, FID signals from a 1:1 complex of BDPA and benzene. The intermediate frequency (IF = 500 MHz) signals in the heterodyne detection system were recorded by a fast digitizer to observe the FID. The spectrum was taken at room temperature with a single pulse and single scan. The receiver is turned on at 80 ns. Top inset, magnified view of boxed area (same parameters and units on axes). Bottom inset, a 0.3-mT-wide c.w. EPR spectrum of BDPA. **b**, Rabi oscillations of BDPA acquired by Fourier-transform (FT) EPR at room temperature. The intensity of the Fourier-transform EPR signals was measured as a function of pulse length. The damping and asymmetry of the observed oscillations is due to an inhomogeneous distribution of B_1 intensity over the sample. Each point represents a measurement from a single pulse. The differences between measurements 1 and 2 result from pulse-to-pulse fluctuations in the instrument.

described in Fig. 3b. Therefore three EPR peaks are observed from the N electron spins in a single crystal of diamond with the application of magnetic fields along the (100) direction. Figure 3c shows Fourier-transform EPR measurements performed on N spins at 295 K at different magnetic fields. We applied a single FEL pulse with 20-ns duration to perform Fourier-transform EPR. As shown in Fig. 3c, the observed three N EPR peaks were shifted by varying the magnetic field. The separation between neighbouring peaks is 94 MHz, the expected hyperfine splitting for ^{14}N impurities in diamond. Figure 3c shows that the Fourier transform EPR spectrum acquired by the FEL-powered pulsed EPR spectrometer is clearly observable over more than 200 MHz bandwidth. This wide excitation bandwidth is necessary to manipulate all three N spin EPR transitions simultaneously, and is more than 50 times broader than the 3-MHz bandwidth when we employ a lower power solid-state source.

Although single-pulse experiments probe a wealth of information, magnetic resonance regularly utilizes multiple coherent pulses to perform more advanced experiments. The simplest example is the powerful Hahn echo sequence, which uses a second pulse to rotate the spins by 180° and refocuses the spins, eliminating the effects of static inhomogeneities in the local fields. Thus, echo experiments can allow for the detection of EPR signals in samples where inhomogeneous broadening causes a rapidly decaying FID, and can also be used to determine the spin decoherence time T_2 of a system. Echo signals were acquired on 4-amino-2,2,6,6-tetramethylpiperidine-1-oxyl (TEMPO) stable radicals dissolved in a deuterated glass

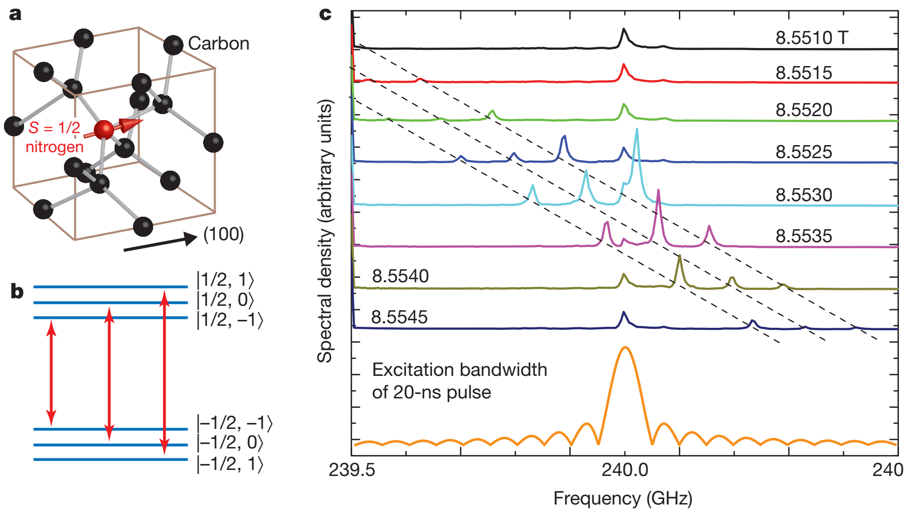


Figure 3 | Fourier-transform EPR measurements with diamond. a, A single nitrogen impurity in a diamond lattice. The nitrogen has an unpaired electron spin ($S = 1/2$). **b**, Owing to hyperfine couplings to the ^{14}N nuclear spin, the energy states of the nitrogen electron spin are split into six states (blue) in a high magnetic field and three pronounced EPR lines are observed, corresponding to the red transitions shown. The spin states are represented by $|m_S, m_I\rangle$, where m_S and m_I are magnetic quantum numbers of electron and nuclear spins respectively. **c**, Fourier-transform EPR of single nitrogen impurities in diamond taken at different magnetic fields (shown on traces). The diamond crystal contains $\sim10^{15}$–10^{16} N spins. The spectra were taken at room temperature, each being an average of 60 traces. The graph shows three EPR peaks because of hyperfine couplings between N electron and ^{14}N nuclear spins (see **b**). The peaks at 240 GHz, which are independent of the strength of the magnetic field, are due to leakage of the FEL radiation. The calculated excitation bandwidth of a square 20-ns pulse is shown at the bottom of the figure.

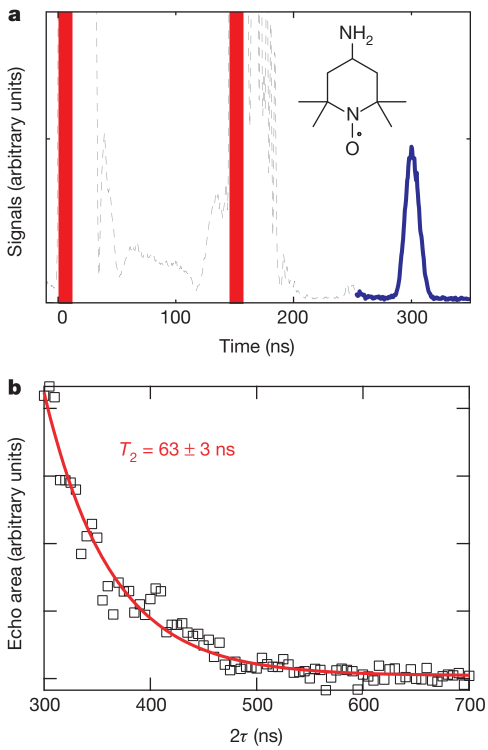


Figure 4 | Hahn echo measurements with TEMPO. a, Hahn echo sequence of a nitroxide spin label (TEMPO, dissolved in a deuterated glass; see main text) showing two 10-ns sliced pulses (represented by red bars) followed by the echo signal (represented by a blue line) taken at 190 K with 7 averages. The receiver is turned on at 250 ns. Signal before that (dashed line) is from incomplete isolation. Inset, the TEMPO molecule. **b**, The echo area (the area under the blue line in Fig. 4a) is plotted as a function of 2τ, where τ is the interpulse spacing, showing an echo decay over several hundred nanoseconds. The red line shows a fit to an exponential decay with T_2 = 63 ns.

composed of a frozen mixture of deuterated glycerol and D_2O (the volume ratio of 60:40). TEMPO is often used as a 'spin label' in EPR studies of the structure and dynamics of biological molecules. The sample was roughly 10 μl of a frozen TEMPO solution of 50 mM radical concentration in a Teflon bucket. Measurements were carried out at from 120 K to 190 K. Figure 4a shows the two pulses used to generate the echo, and the echo signal itself for a single inter-pulse spacing of 150 ns at 190 K. A traditional experiment to determine the spin decoherence time was carried out by recording the decay of the echo area with increasing inter-pulse spacing (Fig. 4b). The pulses used are only ~10 ns long. These short pulses make dramatically shorter relaxation times accessible, increase the bandwidth, and improve the signal-to-noise ratio by exciting more spins than the longer and weaker pulses available from a solid-state source. The single-exponential spin echo decay measured at 190 K is shown in Fig. 4b. The decay time T_2 at 120 K is 225 ± 8 ns. At 190 K, T_2 equals 63 ± 3 ns, an order of magnitude faster than what can be resolved for a spin-1/2 electron with our EPR spectrometer using a solid-state source. This enables T_2 measurements of nitroxide samples to be carried out at 190 K, nearly 100 K higher than is possible with a low-power solid-state source, and very close to the 'glass transition' above which proteins become flexible and are functional.

Currently, sensitivity of the 190 K nitroxide T_2 measurement is limited primarily by the facts that (1) the minimum inter-pulse spacing, or dead time, is 150 ns (2τ = 300 ns), much longer than T_2, and (2) the excitation bandwidth (~100 MHz for 10-ns excitation pulses) is much smaller than the width of the nitroxide spectrum at 8.5 T. Neither of these limitations is fundamental. To reduce the dead time to be comparable to the duration of our ~10-ns excitation pulses, we are minimizing spurious reflections in corrugated waveguides and quasi-optics, and improving our induction-mode isolator following methods outlined in ref. 18. To increase the excitation bandwidth, one requires shorter pulses and thus higher power reaching the sample. We are in the process of constructing a new FEL that is optimized for pulsed EPR at 8.5 T and above, and which should enable ~1-ns excitation pulses (~1-GHz excitation bandwidth). Together, decreasing dead time to 10 ns and broadening the excitation bandwidth to 1 GHz would increase the concentration sensitivity of the nitroxide T_2 measurements presented here by three orders of magnitude.

We are developing methods to control the relative phase of a sequence of FEL pulses at 240 GHz, allowing for signal averaging and pulse/phase cycling, as well as methods for generating more sophisticated pulse sequences. FELs become more powerful and easier to use with increasing frequency, and the pulse slicing technology demonstrated here is scalable to much higher frequencies. Our spectrometer thus shows that pulse-sliced FELs can break through the bottleneck that has thus far limited high-power pulsed EPR to frequencies below 100 GHz. There are no obvious technological barriers to a FEL-based pulsed EPR spectrometer operating above 1 THz, thus taking advantage of the highest static magnetic fields available to realize the full potential of EPR spectroscopy.

METHODS SUMMARY

For c.w. and low-power pulsed EPR measurements and the injection-locking, we employ a 30-mW solid-state source at 240 GHz available from Virginia Diodes, Inc. (VDI). The c.w. EPR measurements were performed by varying the magnetic field of the sweep coil in the 12.5-T superconducting magnet while the main coil of the 12.5-T magnet was in persistent mode. The power of the VDI source was attenuated with a pair of wire-grid polarizers to optimize signal-to-noise ratio of the EPR signals.

For FEL-powered pulsed EPR, the FEL was injection-locked by the 30-mW solid-state source. The injection-locking system realizes extreme phase-stability of the FEL. The linewidth of the FEL is less than 1 MHz which is Fourier-transform-limited by a few-microsecond FEL pulse[22]. Therefore the inter-pulse spacing for the spin echo sequence is limited by the length of the FEL pulse instead of phase stability.

BDPA and TEMPO were purchased from Sigma-Aldrich. Single crystals of type-Ib diamond were obtained from Sumitomo Electric.

Received 18 April; accepted 24 July 2012.

1. Hubbell, W. L., Mchaourab, H. S., Altenbach, C. & Lietzow, M. A. Watching proteins move using site-directed spin labeling. *Structure* **4,** 779–783 (1996).
2. Pannier, M., Veit, S., Godt, A., Jeschke, G. & Spiess, H. W. Dead-time free measurement of dipole-dipole interactions between electron spins. *J. Magn. Reson.* **142,** 331–340 (2000).
3. Saxena, S. & Freed, J. H. Double quantum two dimensional Fourier transform electron spin resonance: distance measurements. *Chem. Phys. Lett.* **251,** 102–110 (1996).
4. Borbat, P. & Freed, J. H. Multiple-quantum ESR and distance measurements. *Chem. Phys. Lett.* **313,** 145–154 (1999).
5. Smilowitz, L. *et al.* Photoexcitation spectroscopy of conducting polymer-C_{60} composites: photoinduced electron transfer. *Phys. Rev. B* **47,** 13835–13842 (1993).
6. Dyakonov, V. *et al.* Photoinduced charge carriers in conjugated polymer-fullerene composites studied with light-induced electron-spin resonance. *Phys. Rev. B* **59,** 8019–8025 (1999).
7. Ogiwara, T., Ikoma, T., Akiyama, K. & Tero-Kubota, S. Spin dynamics of carrier generation in a photoconductive C_{60}-doped poly(N-vinylcarbazole) film. *Chem. Phys. Lett.* **411,** 378–383 (2005).
8. McCamey, D. R. *et al.* Spin Rabi flopping in the photocurrent of a polymer light-emitting diode. *Nature Mater.* **7,** 723–728 (2008).
9. Takahashi, S., Hanson, R., van Tol, J., Sherwin, M. S. & Awschalom, D. D. Quenching spin decoherence in diamond through spin bath polarization. *Phys. Rev. Lett.* **101,** 047601 (2008).
10. Gruber, A. *et al.* Scanning confocal optical microscopy and magnetic resonance on single defect centers. *Science* **276,** 2012–2014 (1997).
11. Lyon, S. A. Spin-based quantum computing using electrons on liquid helium. *Phys. Rev. A* **74,** 052338 (2006).
12. Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. *Rev. Mod. Phys.* **79,** 1217–1265 (2007).
13. Takahashi, S. *et al.* Decoherence in crystals of quantum molecular magnets. *Nature* **476,** 76–79 (2011).
14. Freed, J. H. New technologies in electron spin resonance. *Annu. Rev. Phys. Chem.* **51,** 655–689 (2000).

15. Earle, K. A., Dzikovski, B., Hofbauer, W., Moscicki, J. K. & Freed, J. H. High-frequency ESR at ACERT. *Magn. Reson. Chem.* **43,** S256–S266 (2005).
16. van Tol, J. *et al.* High-field phenomena of qubits. *Appl. Magn. Reson.* **36,** 259–268 (2009).
17. Hofbauer, W., Earle, K. A., Dunnam, C. R., Moscicki, J. K. & Freed, J. H. High-power 95 GHz pulsed electron spin resonance spectrometer. *Rev. Sci. Instrum.* **75,** 1194–1208 (2004).
18. Cruickshank, P. A. S. *et al.* A kilowatt pulsed 94 GHz electron paramagnetic resonance spectrometer. with high concentration sensitivity, high instantaneous bandwidth, and low dead time. *Rev. Sci. Instrum.* **80,** 103102 (2009).
19. Communications and Power Industries (CPI) http://www.cpii.com/product.cfm/7/40.
20. Nanni, E. A., Shapiro, M. A., Sirigiri, J. R. & Temkin, R. J. Design of a 250 GHz photonic band gap gyrotron amplifier. *2010 IEEE Int. Vacuum Electron. Conf.* http://dx.doi.org/10.1109/IVELEC.2010.5503423 (IEEE, 2010).
21. Ramian, G. the new UCSB free-electron lasers. *Nucl. Instrum. Methods Phys. A* **318,** 225–229 (1992).
22. Takahashi, S., Ramian, G., Sherwin, M. S., Brunel, L.-C. & van Tol, J. Submegahertz linewidth at 240 GHz from an injection-locked free-electron laser. *Appl. Phys. Lett.* **91,** 174102 (2007).
23. Takahashi, S., Ramian, G. & Sherwin, M. S. Cavity dumping of an injection-locked free-electron laser. *Appl. Phys. Lett.* **95,** 234102 (2009).
24. Hegmann, F. A. & Sherwin, M. S. Generation of picosecond far-infrared pulses using laser-activated semiconductor reflection switches. *Proc. SPIE* **2842,** 90–105 (1996).
25. Doty, M. F., Cole, B. E., King, B. T. & Sherwin, M. S. Wavelength-specific laser-activated switches for improved contrast ratio in generation of short THz pulses. *Rev. Sci. Instrum.* **75,** 2921–2925 (2004).
26. Rolland, C. & Corkum, P. B. Generation of 130-fsec midinfrared pulses. *J. Opt. Soc. Am. B* **3,** 1625–1629 (1986).
27. Mitsudo, S. *et al.* Development of a sub-THz cw gyrotron for the millimeter wave pulsed ESR spectrometer. *Proc. IRMMW-THz 2010* http://dx.doi.org/10.1109/ICIMW.2010.5612381 (IEEE, 2010).
28. Smith, G. M., Lesurf, J. C. G., Mitchell, R. H. & Riedi, P. C. Quasi-optical cw mm-wave electron spin resonance spectrometer. *Rev. Sci. Instrum.* **69,** 3924–3937 (1998).
29. van Tol, J., Brunel, L.-C. & Wylde, R. J. A quasioptical transient electron spin resonance spectrometer operating at 120 and 240 GHz. *Rev. Sci. Instrum.* **76,** 074101 (2005).
30. Rabi, I. I. Space quantization in a gyrating magnetic field. *Phys. Rev.* **51,** 652–654 (1937).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the NSF (CHE-0821589, DMR-0520481 and DMR-0703925) and the W. M. Keck Foundation. We thank D. Enyeart, S. El Abbadi, N. Krauss, K. Akabori, M. Anholm, T. Visher, J. Bricker and G. Kontsevich for support of the development and operation of the FEL.

Author Contributions S.T. and M.S.S. contributed to the writing of the manuscript. M.S.S., S.H., L.-C.B. and J.v.T. conceived the development of the FEL-powered EPR spectrometer. The development was carried out by S.T., D.T.E., G.R., L.-C.B., S.H. and M.S.S. S.T., D.T.E., J.v.T. and M.S.S. conceived the EPR experiments. The measurements were carried out by S.T., D.T.E. and L.-C.B.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S.S. (sherwin@physics.ucsb.edu).

doi:10.1038/nature11439

High-performance bulk thermoelectrics with all-scale hierarchical architectures

Kanishka Biswas[1]†, Jiaqing He[1,2]†, Ivan D. Blum[2], Chun-I Wu[3], Timothy P. Hogan[3], David N. Seidman[2], Vinayak P. Dravid[2] & Mercouri G. Kanatzidis[1,4]

With about two-thirds of all used energy being lost as waste heat, there is a compelling need for high-performance thermoelectric materials that can directly and reversibly convert heat to electrical energy. However, the practical realization of thermoelectric materials is limited by their hitherto low figure of merit, *ZT*, which governs the Carnot efficiency according to the second law of thermodynamics. The recent successful strategy of nanostructuring to reduce thermal conductivity has achieved record-high *ZT* values in the range 1.5–1.8 at 750–900 kelvin[1–3], but still falls short of the generally desired threshold value of 2. Nanostructures in bulk thermoelectrics allow effective phonon scattering of a significant portion of the phonon spectrum, but phonons with long mean free paths remain largely unaffected. Here we show that heat-carrying phonons with long mean free paths can be scattered by controlling and fine-tuning the mesoscale architecture of nanostructured thermoelectric materials. Thus, by considering sources of scattering on all relevant length scales in a hierarchical fashion—from atomic-scale lattice disorder and nanoscale endotaxial precipitates to mesoscale grain boundaries—we achieve the maximum reduction in lattice thermal conductivity and a large enhancement in the thermoelectric performance of PbTe. By taking such a panoscopic approach to the scattering of heat-carrying phonons across integrated length scales, we go beyond nanostructuring and demonstrate a *ZT* value of ~2.2 at 915 kelvin in p-type PbTe endotaxially nanostructured with SrTe at a concentration of 4 mole per cent and mesostructured with powder processing and spark plasma sintering. This increase in *ZT* beyond the threshold of 2 highlights the role of, and need for, multiscale hierarchical architecture in controlling phonon scattering in bulk thermoelectrics, and offers a realistic prospect of the recovery of a significant portion of waste heat.

The performance of a thermoelectric material is quantified by $ZT = \sigma S^2/(\kappa_{el} + \kappa_{lat})$, where σ is the electrical conductivity, S is the Seebeck coefficient, T is the temperature, κ_{el} is the electronic thermal conductivity and κ_{lat} is the lattice thermal conductivity[4,5]. Among the high-*ZT* materials, PbTe (refs 1–3) is the most efficient for power-generation applications at high temperature, whereas Bi_2Te_3-based materials[6,7] are renowned for refrigeration near room temperature. Several innovative strategies have recently been introduced to increase the *ZT* value of PbTe (refs 1–3). Nanostructuring, in particular, has been proven to be an effective approach to enhance *ZT* by reducing κ_{lat} through the placement of suitable nanoscale precipitates in the matrix, for example in $AgPb_mSbTe_{m+2}$ (ref. 8; LAST), $NaPb_xSbTe_{2+x}$ (ref. 9; SALT) and PbTe–PbS (ref. 10). Alternatively, p-type $PbTe_{1-x}Se_x$ (ref. 11) and Tl–PbTe (ref. 12) also have excellent thermoelectric properties, arising from multiple valence bands and the introduction of a density-of-states distortion in the valence band, respectively. Skutterudite structures have also been shown to have high *ZT* values[13]. Yet, despite remarkable progress, all state-of-the-art materials have *ZT* values in the range of 1.5–1.8 at 750–900 K, well below the target of 2 sought in the field.

Optimized atomic-scale doping/substitution (Fig. 1a) in the PbTe structure can lead to a *ZT* value of ~1.1 at 775 K (Fig. 1b) in the case of a bulk ingot sample of PbTe doped with 2 mol% Na. The maximum

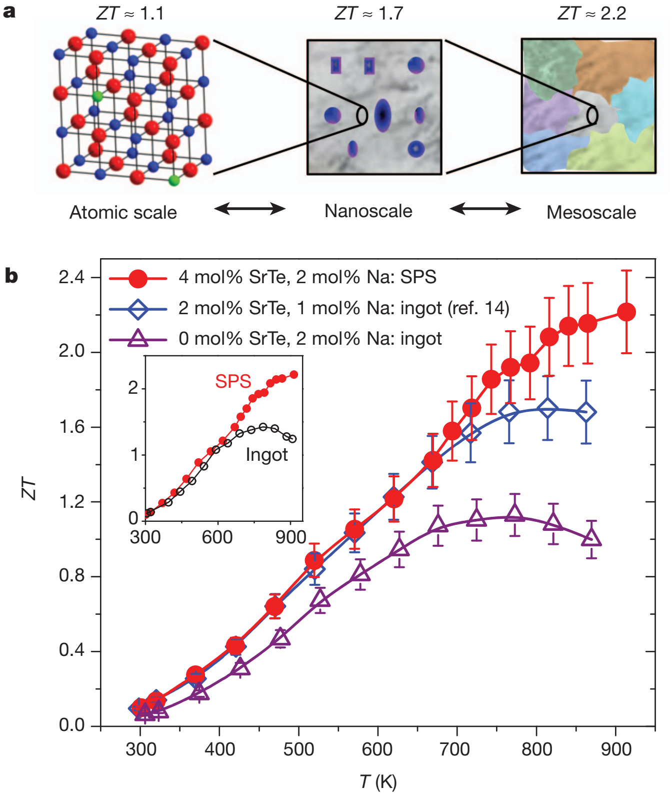


Figure 1 | All-length-scale hierarchy in thermoelectric materials. **a**, Maximum achievable *ZT* values for the respective length scales: the atomic scale (alloy scattering: red, Te; blue, Pb; green, dopant), the nanoscale (PbTe matrix, grey; SrTe nanocrystals, blue) to the mesoscale (grain-boundary scattering). By combining the effects of atomic-scale alloy doping, endotaxial nanostructuring and mesoscale grain-boundary control, maximum phonon scattering can be achieved at high temperatures and the figure of merit can be increased beyond the value possible with nanostructuring alone. **b**, *ZT* as a function of temperature for an ingot of PbTe doped with 2 mol% Na (atomic scale), PbTe–SrTe(2 mol%) doped with 1 mol% Na (ref. 14; atomic plus nanoscale) and spark-plasma-sintered PbTe–SrTe(4 mol%) doped with 2% Na (atomic plus nano plus mesoscale). The measurement uncertainty of all experimental *ZT* versus *T* data was 10% (error bars). Inset, comparison of *ZT* in SPS and ingot samples with the same composition (PbTe–SrTe(4 mol%) doped with 2 mol% Na).

[1] Department of Chemistry, Northwestern University, Evanston, Illinois 60208, USA. [2] Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA. [3] Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan 48824, USA. [4] Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. †Present addresses: New Chemistry Unit, Jawarharlal Nehru Centre for Advanced Scientific Research (JNCASR), Jakkur, Bangalore 560064, India (K.B.); Frantier Institute of Science and Technology (FIST), Xi'an Jiaotong University, Xi'an 710054, China (J.H.).

ZT value increases to ~1.7 at 800 K (Fig. 1b) on the introduction of 2–10-nm endotaxial SrTe nanocrystals (Fig. 1a) into the Na-doped PbTe matrix[14]. This performance increase stems from the fact that the nanostructures impede much of the heat flow in this system while leaving the hole mobility largely unaffected[14]. Nanostructuring itself, however, scatters phonons with short and medium mean free paths (~3–100 nm), thus rendering only phonons with longer mean free paths unaffected. An additional and significant reduction in κ_{lat} may be achieved by further scattering of the phonons with longer mean free paths (~0.1–1 μm, that is, mesoscale; Fig. 1a), on which scale additional mechanisms of grain-boundary phonon scattering and impedance can be exploited. Grain-boundary phonon scattering has been shown to be important in improving the thermoelectric performance of Bi_2Te_3-based alloys, PbTe and nanostructured silicon[6,15,16].

Here we go beyond nanostructuring and show that by harnessing integrated phonon scattering across multiple length scales—atomic-scale alloy scattering, scattering from nanoscale endotaxial precipitation and scattering from mesoscale grain boundaries—we can achieve a record-high *ZT* value of ~2.2 at 915 K (Fig. 1b) in powder-processed and spark-plasma-sintered (SPS) samples of PbTe–SrTe(4 mol%) doped with 2 mol% Na. Compared with melt-processed ingot specimens of the same composition (Fig. 1b, inset), SPS specimens show a ~30–50% increase in *ZT*, underscoring the role of integrated all-length-scale scattering of heat-carrying phonons as reflected in a reduced κ_{lat} value. This represents a panoscopic approach (a hierarchical architecture across all relevant length scales) to tackling the challenge of increasing *ZT*, and is an excellent example of a bulk system in which all relevant length scales are harnessed for achieving effective and efficient phonon scattering and, consequently, a record-high value of *ZT*. This approach is applicable to all bulk thermoelectric materials.

Samples of composition PbTe–SrTe(0–4 mol%) doped *p*-type with 2 mol% Na were synthesized in ingot form by melting and quenching, followed by powder processing and spark plasma sintering into dense pellets (>98% of the theoretical density). The temperature-dependent

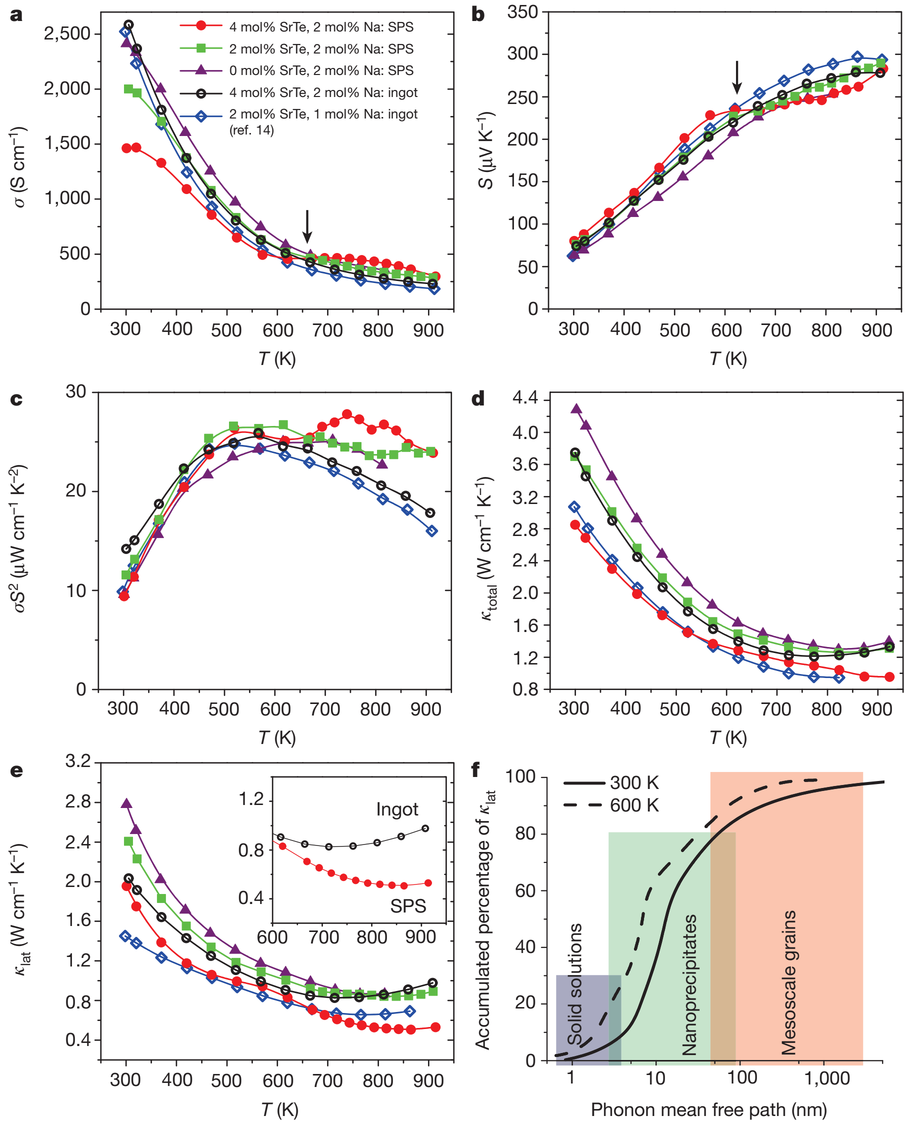


Figure 2 | Thermoelectric properties of SPS and ingot samples of PbTe–SrTe doped with 2 mol% Na. a, Temperature-dependent electrical conductivity (*σ*). The same symbol notation for the samples is used in all panels. The temperature-dependent transport data for the bulk ingot of PbTe–SrTe(2 mol%) doped with 1 mol% Na (ref. 14) is shown for comparison. **b**–**e**, Temperature-dependent Seebeck coefficient (*S*; **b**), power factor (σS^2; **c**), total thermal conductivity (κ_{total}; **d**) and lattice thermal conductivity (κ_{lat}; **e**). Inset in **e**, comparison of κ_{lat} in SPS and ingot samples with the same composition (PbTe–SrTe(4 mol%) doped with 2 mol% Na). **f**, Contributions of phonons with different mean free paths to the cumulative κ_{lat} value for PbTe, adapted from the literature[23]. Phonons with short, medium and long mean free paths can be scattered by atomic-scale defects, nanoscale precipitates and mesoscale grain boundaries, respectively. Black arrows in **a** and **b** indicate the temperature of onset of Na diffusion in Na-doped PbTe, which results in the observed plateaux, starting at ~650 K, in *σ* and *S* as functions of *T*.

electrical and thermal transport properties of several such samples in the 300–950 K range are given in Fig. 2. The transport measurements of SPS samples were compared to the corresponding bulk ingot samples, and also with a previously reported ingot of PbTe–SrTe(2 mol%) doped with 1 mol% Na (ref. 14). Typically, the 2 mol% Na doped SPS sample containing 4 mol% SrTe has a electrical conductivity of $\sigma \approx 1{,}465\ \mathrm{S\,cm^{-1}}$ at room temperature, which decreases to $\sim 457\ \mathrm{S\,cm^{-1}}$ at 620 K, remains almost at that value in the 620–850 K range and reaches a value of $\sim 300\ \mathrm{S\,cm^{-1}}$ at ~915 K (Fig. 2a). Ingot samples with the same nominal composition (PbTe–SrTe(4 mol%) doped with 2 mol% Na) show the room-temperature value of $\sigma \approx 2{,}585\ \mathrm{S\,cm^{-1}}$, which decreases to $\sim 230\ \mathrm{S\,cm^{-1}}$ at 910 K (Fig. 2a). The σ values of the SPS samples are higher at high temperatures, and effectively reach a plateau from 600 to 850 K. This causes the power factor, σS^2, to remain high at higher temperatures. We attribute this rise in σ at high temperature to the enhanced dissolution in the PbTe matrix of the Na dopant, which at lower temperatures remains confined and segregated at grain boundaries, following powder and spark plasma sintering. This enhanced dissolution lowers the Fermi level further and generates new charge-carrier holes particularly in the 'heavy-hole' valence band of PbTe (refs 17–19). We confirmed the proposed segregation of Na at grain boundaries, precipitate–matrix interfaces and dislocations with three-dimensional atom probe tomography (APT; see below).

The Seebeck coefficient, S, is in agreement with the Hall measurements (Supplementary Information) for p-type PbTe (Fig. 2b). Typically, at room temperature the 4 mol% SrTe SPS sample doped with 2% Na has $S \approx 81\ \mathrm{\mu V\,K^{-1}}$, which rapidly increases with temperature, remains almost constant in the 600–850 K range and then reaches a value $\sim 284\ \mathrm{\mu V\,K^{-1}}$ at ~915 K (Fig. 2b). The high values of S at high temperature arise from the well-known contribution of the two valence bands in PbTe (refs 14, 17–19).

Figure 2c shows the power factors of different SPS and ingot PbTe–SrTe samples doped with 2% Na, as functions of temperature. Compared with ingots, the SPS samples have higher σS^2 values at higher temperatures, 600–915 K, because of the higher σ values in this range. Typically, the room-temperature σS^2 value we measured was $\sim 9.5\ \mathrm{\mu W\,cm^{-1}\,K^{-2}}$ for the 4 mol% SrTe SPS sample doped with 2% Na. This value rose to a maximum ($\sim 28\ \mathrm{\mu W\,cm^{-1}\,K^{-2}}$) at about 745 K and fell to $\sim 24\ \mathrm{\mu W\,cm^{-1}\,K^{-2}}$ at ~915 K.

The total thermal conductivity, κ_{total}, of the various samples is shown in Fig. 2d. A typical room-temperature value of $\kappa_{\mathrm{total}} \approx 2.85\ \mathrm{W\,m^{-1}\,K^{-1}}$ was observed for PbTe–SrTe(4 mol%) SPS samples doped with 2% Na, which decreased to $\sim 0.96\ \mathrm{W\,m^{-1}\,K^{-1}}$ at ~923 K. The electrical thermal conductivity ($\kappa_{\mathrm{el}} = L\sigma T$, where L is the Lorenz number) was estimated on the basis of an L value obtained from the accepted approach of fitting the Seebeck data to the reduced chemical potential[10,20,21] (Supplementary Fig. 6). The lattice thermal conductivity, κ_{lat}, was estimated by subtracting κ_{el} from κ_{total} (Fig. 2e). We observe that the κ_{lat} values of the SPS samples are lower than those of the ingots with the corresponding compositions and, more importantly, that the difference increases as the temperature is increased (Fig. 2e, inset), which leads to significantly depressed values at ~900 K compared with the ingots. The upturn in the plot of κ_{lat} as a function of T at high temperature for the ingot sample is due to the bipolar contribution to the thermal conductivity by thermally generated electrons and holes[22]. In contrast, in the SPS samples the bipolar contribution is negligible and there is no upturn because of the existence of an interfacial potential at grain boundaries that scatters more electrons than holes[6]. In addition, the overall decrease in the κ_{lat} value of the SPS sample compared with the corresponding ingot can be attributed to the additional scattering from and impedance due to mesoscale grain boundaries in the SPS sample, which is important in reducing bipolar conduction and scattering phonons with longer mean free paths, which generally are largely ignore nanostructuring.

The contributions to κ_{lat} of phonons with different mean free paths have recently been calculated for PbTe (refs 23, 24; Fig. 2f), $\mathrm{PbTe}_{1-x}\mathrm{Se}_x$ (ref. 24) and Si (ref. 25). Around ~80% of the κ_{lat} value of PbTe is contributed by phonon modes with mean free paths of less than 100 nm, which can be attributed to scattering by a combination of atomic-scale solid-solution alloying, nanoscale precipitates embedded in PbTe and associated spatially distributed strain[23] (Fig. 2f). The remaining ~20% of κ_{lat} in PbTe, however, is contributed by phonon modes with mean free paths of 0.1–1 μm. The mesoscale grain structure, achieved by spark plasma sintering, is comparable in size to the mean free path and thus can scatter a notable fraction of these additional heat-carrying phonons. This results in further reduction of κ_{lat}, compared with nanostructuring alone.

The SPS samples contain nanoscale precipitates and mesoscale grains and associated grain boundaries, which are clearly evident in transmission electron microscopy (TEM) and APT studies. Detailed microstructure investigations using TEM were carried out on the SPS PbTe–SrTe(4 mol%) sample doped with 2 mol% Na. Typical low- and middle-magnification TEM images are shown in Fig. 3a and Fig. 3b, respectively. The presence of mesoscale grains 0.1–1.7 μm in size and nanoscale precipitates with dark contrasts in the range of 1–17 nm is evident in these images. The nanoscale precipitates have two typical shapes, platelet-like and spherical/ellipsoidal, with three crystallographic variants consistent with bicrystal symmetry. The small precipitates (~1–6 nm) have a platelet-like morphology and are coherently strained, whereas the larger precipitates (~10–17 nm) have spherical or ellipsoidal shapes along with interfacial misfit dislocations. The latter arise from excess coherency strain that derives from the small lattice parameter mismatch (6.460 Å versus 6.660 Å for PbTe and SrTe,

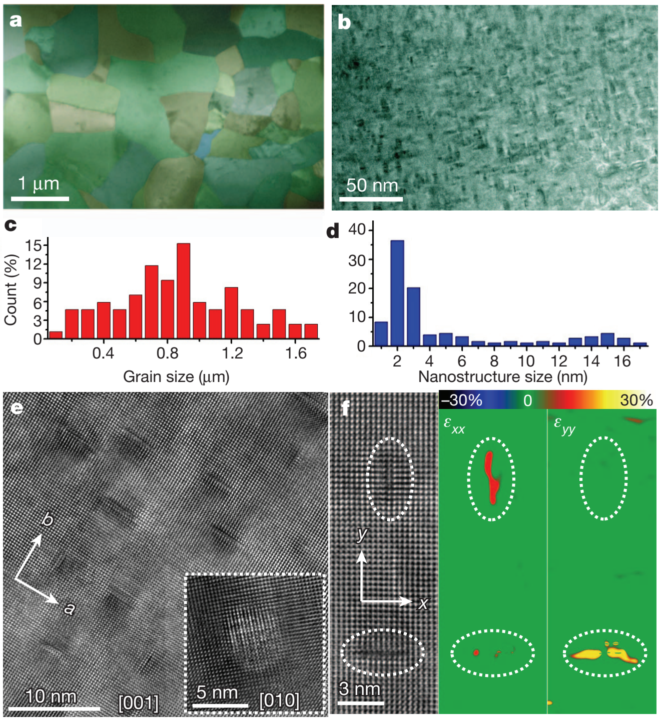


Figure 3 | Micro and nanostructures in SPS PbTe–SrTe(4 mol%) doped with 2 mol% Na. a, Low-magnification TEM image showing mesoscale grains in the sample. **b**, Medium-magnification TEM image revealing presence of platelet-like and spherical/ellipsoidal nanoscale precipitates. **c**, Grain size distribution histogram. **d**, Size distribution histogram of SrTe nanoparticles. **e**, High-magnification lattice image depicting some perpendicular or parallel platelet-like precipitates. Inset, a small spherical precipitate with a coherent interface with the matrix. **f**, Lattice image and strain maps showing elastic strain (colour scale) along only one direction for platelet-like precipitates. The estimate of the distribution density of all types of nanoscale precipitates is $\sim 1.2 \times 10^{12}\ \mathrm{cm^{-2}}$.

respectively). Figures 3c and 3d show the size distribution histogram of the mesoscale grains and nanoscale precipitates, respectively: the average size of the mesoscale grains is ~0.8 µm and that of the nanoscale precipitates is ~2.8 nm.

Figure 3e shows a representative high-resolution TEM image of platelet-like precipitates obtained with the electron beam parallel to the [001] axis. All platelet-like precipitates are organized perpendicular or parallel to each other, consistent with two of three possible crystallographic variants. The inset image, taken from a different region, shows a small spherical precipitate with a coherent (elastically strained) interface with the matrix, without any interfacial dislocations. To analyse the possible strain at and near the precipitate–matrix interface, the high-quality, high-resolution TEM images were analysed by geometric phase analysis[26], which is a semi-quantitative lattice image-processing approach for revealing spatially distributed strain fields. Figure 3f shows the image and the results of the analysis, namely the components ε_{xx} and ε_{yy} of the strain. The image shows two perpendicular platelet-like precipitates enclosed by dotted lines, and indicates that there is elastic strain only along the x direction for the upper precipitate and only along the y direction for the lower. Thus, the strain distribution in platelet-like precipitates is anisotropic, in contrast to spherical precipitates, which have more-uniform omnidirectional strain distributions.

Scanning TEM investigations (Fig. 4a) show the presence of some medium-size (20–50-nm) precipitates, in addition to smaller ones (1–15 nm). Energy dispersion X-ray spectroscopy indicates a large increase in the Sr signal from the precipitates (Fig. 4a, black curve in inset) compared to the matrix regions (Fig. 4a, blue curve in inset), suggesting that they are mainly SrTe.

The presence of SrTe nanoscale precipitates in the PbTe matrix was confirmed independently by APT[27]. The three-dimensional reconstruction of the volume of the sample of PbTe–SrTe(4 mol%) doped with 2% Na analysed by APT is given in Fig. 4b. The compositions of the matrix and the precipitates correspond to SrTe and PbTe, respectively (Supplementary Table 3). The composition profile across the interface of the large precipitate is measured using a proximity histogram (Fig. 4c), which shows ~1.6 at% Na accumulation at the interface. There is also a slight Na concentration gradient in the matrix, with a lower concentration near the interface than in the bulk of the matrix. Sodium also accumulates at the core of a linear defect in the same reconstructed volume; that is, there is segregation at the dislocation core. Sodium was also observed to segregate at grain boundaries (Fig. 4d and Supplementary Fig. 8). We believe that the Na which is confined to grain boundaries (and other defect sites) at low temperature goes back into solid solution with the PbTe matrix at elevated temperatures, thus increasing the p-type charge-carrier density. This provides a viable explanation for the enhanced electrical conductivity (and power factor) of SPS samples at high temperature as discussed above (Fig. 2). Fitting the experimental diffusion coefficient, D, to $1/T$ data for Na in PbTe (ref. 28) and extrapolating to lower temperatures yields $D = 1.0 \times 10^{-16}\,\mathrm{cm^2\,s^{-1}}$ and $D = 3.6 \times 10^{-14}\,\mathrm{cm^2\,s^{-1}}$ at 550 and 650 K, respectively. Thus, the root mean square diffusion distance ($\sqrt{4Dt}$, where t is the diffusion time) of Na is ~10 nm for $t = 40$ min at 550 K and for $t = 7$ s at 650 K. This can be considered the temperature range for the onset of Na diffusion in Na-doped PbTe, which results in the observed plateaux, starting at ~650 K, in σ and S as functions of T (Fig. 2a, b, arrows).

The panoscopic approach goes beyond nanostructuring and takes advantage of all relevant length scales by including the effects of mesoscale grain boundaries, endotaxial nanostructuring and atomic-scale substitutional doping in a bulk material. In this way, more extensive phonon scattering can be achieved and thermoelectric performance can be maximized. The p-type PbTe–SrTe system illustrates the important role (at high temperature) of grain-boundary phonon

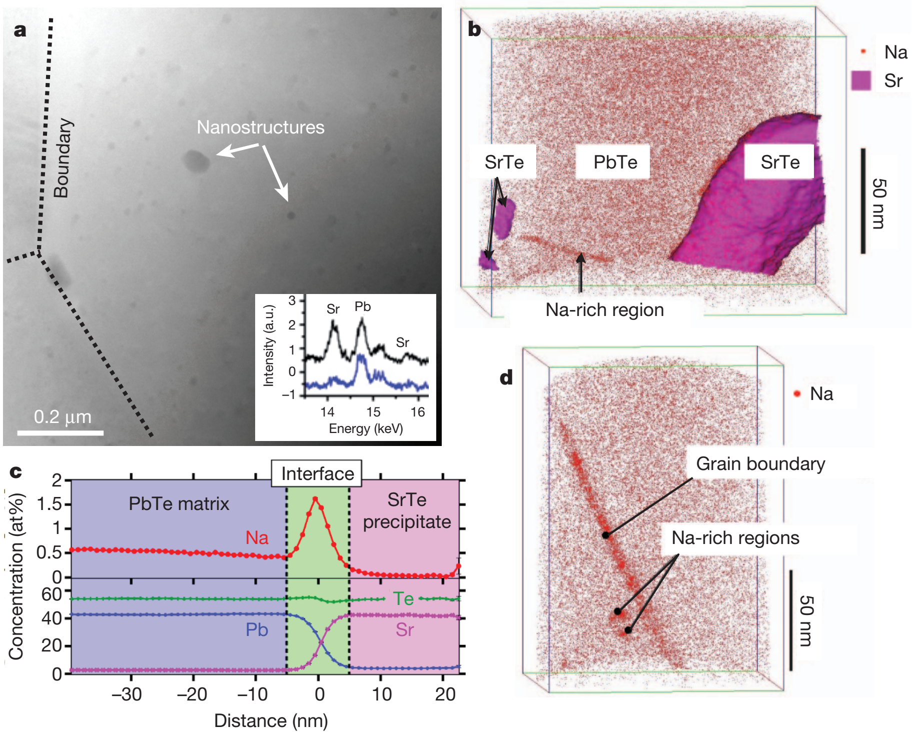


Figure 4 | Compositional analysis of SPS PbTe–SrTe(4 mol%) doped with 2 mol% Na. a, Scanning TEM image showing the presence of SrTe nanostructures in the PbTe matrix. Inset, energy dispersion X-ray spectrum (black, precipitate; blue, matrix). a.u., arbitrary units. **b**, Three-dimensional reconstruction of the volume analysed by APT (for clarity, only half of the Na atoms are displayed). SrTe precipitates are highlighted using a 25 at% Sr isoconcentration surface. **c**, Proximity histogram showing the concentration profiles of Pb, Te, Sr and Na across the interface of the large SrTe precipitate. **d**, Three-dimensional reconstruction of a volume, analysed by APT, containing a grain boundary (for clarity, only the Na atoms are displayed). The volume is viewed from a direction parallel to the grain boundary.

scattering, which, in combination with nanostructuring, decreases κ_{lat} to levels well below those that can be reached by endotaxial nanostructuring alone. This is coupled to the added benefit of carrier generation at elevated temperatures through the dissolution of otherwise grain-boundary-confined Na into the bulk matrix at lower temperatures. Thus, a ZT value of ~2.2 at 915 K is readily and consistently achievable. The hierarchical architecture approach described here is expected to be applicable to any bulk thermoelectric system. With this advance in the maximum figure of merit, we can expect average ZT values of ~1.2 and ~1.7 for non-segmented and segmented thermoelectric devices, respectively ($ZT \approx 1.2$ at 350 K for segmentation with BiSbTe (ref. 6)). Considering a thermoelectric device with a cold-side temperature of 350 K and a hot-side temperature of 950 K, respective waste-heat conversion efficiencies[3,4] of ~16.5% and ~20% are predicted. This may open realistic pathways to broad-based applications in automotive, military and marine waste-heat recovery.

METHODS SUMMARY

Several samples of PbTe–SrTe(0–4 mol%) doped with 2 mol% Na were synthesized first in the form of bulk ingots by melting at 1,323 K over 10 h, quenching to room temperature (297 K), followed by powder processing (Retsch RM200, Retsch GmbH) and spark plasma sintering (SPS 10-4, Thermal Technology LLC) at 823 K for 10 min under an axial pressure of 60 MPa in an argon atmosphere (supplementary, experimental). The σ and S were measured simultaneously in a helium atmosphere at temperatures ranging from room temperature to about 923 K on a ULVAC-RIKO ZEM-3 instrument system. We determined carrier concentrations using measurements of Hall coefficients at room temperature with a home-built system in applied magnetic fields ranging from 0 to 1.25 T. The thermal diffusivity, D, was directly measured in the temperature range 300–923 K by using the laser flash diffusivity method in a commercial Netzsch LFA-457 instrument. The thermal diffusivity was measured along the same direction as was the electrical transport. The heat capacity, C_p, was determined on the basis of previous reported experimental literature for PbTe (refs 11, 29). The total thermal conductivity was calculated using the formula $\kappa_{\mathrm{total}} = DC_p\rho$, where ρ is the sample density, measured by gas pycnometer (Micromeritics AccuPyc 1340).

Received 31 March; accepted 24 July 2012.

1. Snyder, J. G. & Toberer, E. S. Complex thermoelectric materials. *Nature Mater.* **7,** 105–114 (2008).
2. Chen, G., Dresselhaus, M. S., Dresselhaus, G., Fleurial, J. P. & Caillat, T. Recent development in thermoelectric materials. *Int. Mater. Rev.* **48,** 45–66 (2003).
3. Sootsman, J., Chung, D. Y. & Kanatzidis, M. G. New and old concepts in thermoelectric materials. *Angew. Chem. Int. Ed.* **48,** 8616–8639 (2009).
4. Rowe, D. M. *CRC Handbook of Thermoelectrics: Macro to Nano* (CRC/Taylor & Francis, 2006).
5. Tritt, T. M. (ed.) *Recent Trends in Thermoelectric Materials Research I* (Semiconductors and Semimetals Vol. 69, Academic, 2000); *Recent Trends in Thermoelectric Materials Research II* (Semiconductors and Semimetals Vol. 70, Academic, 2000); *Recent Trends in Thermoelectric Materials Research III* (Semiconductors and Semimetals Vol. 71, Academic, 2001).
6. Poudel, B. *et al.* High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys. *Science* **320,** 634–638 (2008).
7. Venkatasubramanian, R., Siivola, E., Colpitts, V. & O'Quinn, B. Thin-film thermoelectric devices with high room-temperature figures of merit. *Nature* **413,** 597–602 (2001).
8. Hsu, K. F. *et al.* Cubic $\mathrm{AgPb}_m\mathrm{SbTe}_{2+m}$: bulk thermoelectric materials with high figure of merit. *Science* **303,** 818–821 (2004).
9. Poudeu, P. F. P. *et al.* High thermoelectric figure of merit and nanostructuring in bulk p-type $\mathrm{Na}_{1-x}\mathrm{Pb}_m\mathrm{Sb}_y\mathrm{Te}_{2+m}$. *Angew. Chem. Int. Ed.* **45,** 3835–3839 (2006).
10. Girard, S. N. *et al.* High performance Na-doped PbTe-PbS thermoelectric materials: electronic density of states modification and shape-controlled nanostructures. *J. Am. Chem. Soc.* **133,** 16588–16597 (2011).
11. Pei, Y. *et al.* Convergence of electronic bands for high-performance bulk thermoelectric. *Nature* **473,** 66–69 (2011).
12. Heremans, J. P. *et al.* Enhancement of thermoelectric efficiency in PbTe by distortion of the electronic density of states. *Science* **321,** 554–557 (2008).
13. Shi, X. *et al.* Multiple-filled skutterudites: high thermoelectric figure of merit through separately optimizing electrical and thermal transport. *J. Am. Chem. Soc.* **133,** 7837–7846 (2011).
14. Biswas, K. *et al.* Strained endotaxial nanostructures with high thermoelectric figure of merit. *Nature Chem.* **3,** 160–166 (2011).
15. Zhu, G. H. *et al.* Increased phonon scattering by nanograins and point defects in nanostructured silicon with a low concentration of germanium. *Phys. Rev. Lett.* **102,** 196803 (2009).
16. Martin, J., Wang, L., Chen, L. & Nolas, G. S. Enhanced Seebeck coefficient through energy-barrier scattering in PbTe nanocomposite. *Phys. Rev. B* **79,** 115311 (2009).
17. Ravich, Y. I., Efimova, B. A. & Smirnov, I. A. *Semiconducting Lead Chalcogenides* Vol. 5 184–192 (Plenum, 1970).
18. Crocker, A. J. & Rogers, L. M. Valence band structure of PbTe. *J. Phys. Colloq.* **29** (C4) 129–132 (1968).
19. Airapetyants, S. V. & Vinogradova, M. N. Durbrovskaya, I. N., Kolomoets, N. V. & Rudnik, I. M. Structure of the valance band of heavily doped lead telluride. *Sov. Phys. Solid State* **8,** 1069–1072 (1966).
20. Johnsen, S. *et al.* Nanostructures boost the thermoelectric performance of PbS. *J. Am. Chem. Soc.* **133,** 3460–3470 (2011).
21. May, A. F., Fleurial, J.-P. & Snyder, G. J. Thermoelectric performance of lanthanum telluride produced via mechanical alloying. *Phys. Rev. B* **78,** 125205 (2008).
22. Goldsmid, H. J. *Thermoelectric Refrigeration* (Plenum, 1964).
23. Qiu, B., Bao, H., Zhang, G., Wu, Y. & Ruan, X. Molecular dynamics simulations of lattice thermal conductivity and spectral phonon mean free path of PbTe: bulk and nanostructures. *Comput. Mater. Sci.* **53,** 278–285 (2012).
24. Tian, Z. *et al.* Phonon conduction in PbSe, PbTe and $\mathrm{PbTe}_{1-x}\mathrm{Se}_x$ from first-principle calculations. *Phys. Rev. B* **85,** 184303 (2012).
25. Esfarjani, K., Chen, G. & Stokes, H. T. Heat transport in silicon from first-principle calculations. *Phys. Rev. B* **84,** 085204 (2011).
26. Hÿtch, M. J., Snoeck, E. & Kilaas, R. Quantitative measurement of displacement and strain fields from HREM micrographs. *Ultramicroscopy* **74,** 131–146 (1998).
27. Seidman, D. N. Three-dimensional atom-probe tomography: advances and applications. *Annu. Rev. Mater. Res.* **37,** 127–158 (2007).
28. Crocker, A. J. & Dorning, B. F. Diffusion of sodium in lead telluride. *J. Phys. Chem. Solids* **29,** 155–161 (1968).
29. Blachnik, R. & Igel, R. Thermodynamic properties of IV–VI compound: lead chalcogenides. *Z. Naturforsch. B* **29,** 625–629 (1974).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Energy Frontier Research Center for Revolutionary Materials for Solid State Energy Conversion, funded by the US Department of Energy, Office of Science, Basic Energy Sciences under award number DE-SC0001054. TEM was performed in the (EPIC) (NIFTI) (Keck-II) facility of the NUANCE Center at Northwestern University. The spark plasma sintering system at Michigan State University is supported by ONR DURIP. APT was performed at NUCAPT and supported by NSF-MRI (DMR-0420532), ONR-DURIP (N00014-0400798, N00014-0610539, N00014-0910781), MRSEC (DMR-1121262) and ISEN at Northwestern University.

Author Contributions K.B. synthesized the samples and designed and carried out thermoelectric experiments. J.H. performed the TEM experiments. I.D.B. performed the APT measurements. C.-I.W. and T.P.H. performed the spark plasma sintering. K.B., J.H., I.D.B., D.N.S., V.P.D. and M.G.K. conceived the experiments, analysed the results and wrote and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.G.K. (m-kanatzidis@northwestern.edu).

doi:10.1038/nature11357

Oceanic nitrogen reservoir regulated by plankton diversity and ocean circulation

Thomas Weber[1] & Curtis Deutsch[1]

The average nitrogen-to-phosphorus ratio of marine phytoplankton (16N:1P) is closely matched to the nutrient content of mean ocean waters (14.3N:1P). This condition is thought to arise from biological control over the ocean's nitrogen budget, in which removal of bioavailable nitrogen by denitrifying bacteria ensures widespread selection for diazotrophic phytoplankton that replenish this essential nutrient when it limits the growth of other species[1–3]. Here we show that in the context of a realistic ocean circulation model, and a uniform N:P ratio of plankton biomass, this feedback mechanism yields an oceanic nitrate deficit more than double its observed value. The critical missing phenomenon is diversity in the metabolic N:P requirement of phytoplankton, which has recently been shown to exhibit large-scale patterns associated with species composition[4]. When we model these variations, such that diazotrophs compete with high N:P communities in subtropical regions, the ocean nitrogen inventory rises and may even exceed the average N:P ratio of plankton. The latter condition, previously considered impossible, is prevented in the modern ocean by shallow circulations that communicate stoichiometric signals from remote biomes dominated by diatoms with low N:P ratios. Large-scale patterns of plankton diversity and the circulation pathways connecting them are thus key factors determining the availability of fixed nitrogen in the ocean.

The biologically mediated feedback between marine denitrification and N_2 fixation operates like a 'nutrient thermostat'[1,2,5] that couples the ocean's fixed N reservoir (primarily in the form of nitrate, NO_3^-) to its less dynamic reservoir of P (primarily phosphate, PO_4^{3-}). A central element of this self-regulating mechanism is the partition of ecological niches between diazotrophic (N_2-fixing) phytoplankton, which grow slowly but do not require an external supply of fixed N (refs 6, 7), and other plankton that grow quickly but are often N-limited owing to persistent N removal in anoxic environments[8–10]. The quantitative understanding of this mechanism rests on box models, in which diazotrophs maintain the ocean's ratio of major nutrient reservoirs ($\Sigma N/\Sigma P$) close to, but slightly below, the N:P requirements of the plankton with which they compete for PO_4^{3-} (refs 5, 11, 12).

Box model depictions make two major simplifications. First, diazotrophs are assumed to compete with plankton having a universal Redfield ratio of 16N:1P, which sets a threshold for N_2 fixation and constitutes the 'set point' of the nutrient thermostat, towards which $\Sigma N/\Sigma P$ is restored. In reality, the Redfield ratio is only an average value and plankton N:P varies systematically between marine species and their preferred biomes[4,13–15]. It has thus been hypothesized that the regulation of $\Sigma N/\Sigma P$ is biased towards the nutrient requirements of those species cohabiting with diazotrophs in subtropical biomes[3]. Second, the circulation pathways that transport nutrients between surface regions with different species composition and N:P ratios are not represented in box models, but may have a central role in shaping the ecological niche of diazotrophs[3]. A realistic physical model is required to identify the processes maintaining the N reservoir of an ocean with diverse plankton stoichiometry.

We developed a simple ecosystem and biogeochemical model to simulate the long-term coupling of N and P cycles in an observationally constrained ocean general circulation model (GCM). The ecosystem comprises a general phytoplankton class (O) that assimilates NO_3^- and PO_4^{3-} in the molar ratio R_O, and a diazotrophic class (F) that assimilates PO_4^{3-} and releases NO_3^- from newly fixed N_2. The maximum growth rate of N_2-fixers (μ_F) is reduced relative to that of other phytoplankton (μ_O), to reflect a constant energetic cost of diazotrophy and a variable dependence on iron (Fe)[16,17]. Fe-limitation is patterned according to the distribution of atmospheric dust deposition, and its overall strength is varied between simulations (Fig. 1a, Supplementary Fig. 1). Denitrification is simulated in benthic grid boxes and anoxic regions of the water column (Supplementary Fig. 2), and its global rate is also varied within specified limits. The model's P reservoir is conserved at the modern ocean value, but its N inventory adjusts over millennial timescales until a steady-state balance between N_2-fixation and denitrification is achieved. See Methods for model details.

Consistent with box models[5], we found that in a 'Redfieldian' ocean where $R_O = 16$ everywhere, the steady-state value of $\Sigma N/\Sigma P$ depends on two parameters: (1) the globally integrated rate of denitrification, and (2) the competitive handicap faced by diazotrophs (μ_F/μ_O). The decrease in $\Sigma N/\Sigma P$ at higher denitrification rates and stronger Fe limitation (Fig. 1b) reflects the proportions of PO_4^{3-} and NO_3^- required to support global N_2 fixation rates that balance N removal through denitrification. Because diazotrophs have slow growth rates, they compete successfully for PO_4^{3-} only when NO_3^- is low enough to hinder their competitors to a similar degree. The diazotrophic niche is thus determined, through competitive dynamics, by the $NO_3^-:PO_4^{3-}$ ratio of ambient sea water, even though their growth is explicitly independent of NO_3^-. When strong Fe limitation exacerbates their competitive handicap, the deep ocean must accumulate a larger deficit of NO_3^- (lower $\Sigma N/\Sigma P$) to support the required N_2-fixation rate. Similarly, higher global rates of denitrification must be balanced by enhanced N_2 fixation that, for a given diazotrophic growth rate, can only be achieved with a greater excess of PO_4^{3-} (lower $\Sigma N/\Sigma P$).

Throughout the observationally supported range of global denitrification rates (150–250 teragrams of N per year, Tg N yr^{-1}, ref. 18), simulated $\Sigma N/\Sigma P$ is considerably lower than its observed value of approximately 14.3. For the average plankton, this amounts to a global deficit of NO_3^- relative to PO_4^- of 6–13 μM, which is 2–4 times larger than observed. Even when Fe limitation is eliminated, unrealistically low denitrification rates (<100 Tg N yr^{-1}) are required to reconcile the model with observations (Fig. 1b). Neither a shift in the patterns of denitrification nor greater model complexity—adding dissolved organic matter, a complete iron cycle, or minor N budget terms—can eliminate this discrepancy (see Supplementary Notes). It can only be resolved by expanding the ecological niche of diazotrophs. This cannot be accomplished by increasing their growth rate, which already reaches near parity with other plankton. It requires a process through which the availability of PO_4^{3-} is enhanced relative to NO_3^- in the subtropics. We investigated whether large-scale deviations from

[1] University of California Los Angeles, Los Angeles, California 90095, USA.

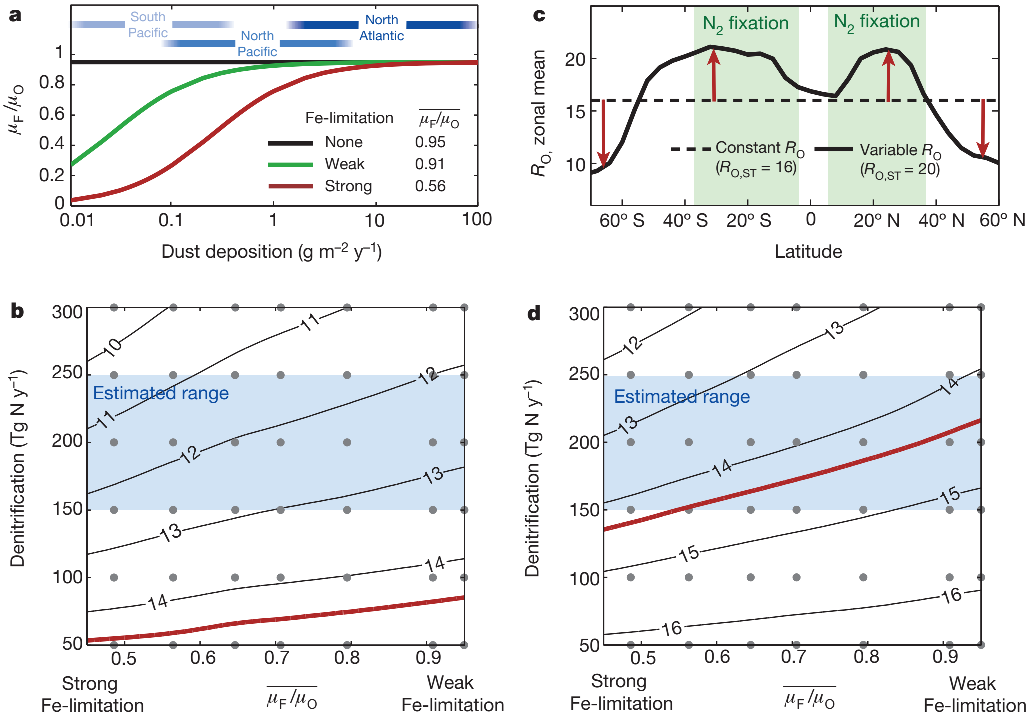


Figure 1 | Model scenarios and solutions. **a**, Growth rate of diazotrophs relative to other phytoplankton, as a function of atmospheric dust deposition. When Fe limitation is stronger, μ_F/μ_O is more variable between regions of high and low deposition, and its mean value ($\overline{\mu_F/\mu_O}$) is reduced. **b**, Predicted steady-state ΣN/ΣP (black contours) for a Redfieldian ocean, across a range of denitrification and Fe-limitation scenarios. Blue shading represents observational constraints on denitrification range; red line indicates observed ΣN/ΣP; grey dots are solutions of individual simulations. **c**, Variations in R_O are incorporated using inferred community composition. Diazotrophs are predominantly confined to the green-shaded latitude bands. **d**, As for **b**, except for stoichiometrically diverse scenarios ($R_{O,ST} \approx 20$).

Redfield stoichiometry in nutrient uptake by non-fixing phytoplankton[3] could provide such a mechanism, and resolve the gap between model predictions and measurements.

Large-scale variations in R_O have recently been shown to hold throughout the Southern Ocean, where polar latitudes dominated by diatoms export low N:P organic matter, while Subantarctic latitudes are characterized by high N:P export ratios[4]. We added stoichiometric diversity to the plankton in our model by extending this empirically derived relationship between plankton biogeography and biomass N:P (see Methods), while ensuring a global mean nutrient export ratio of 16N:1P (Fig. 1c, Supplementary Figs 3 and 4). Because diatoms are abundant in equatorial and high latitudes but scarce in the subtropics, the mean N:P of plankton that compete directly with diazotrophs—denoted $R_{O,ST}$—is close to 20. This is consistent with the elemental composition of the cyanobacteria that dominate oligotrophic waters[19,20], the observed N:P of organic matter in the North Pacific Subtropical Gyre[21], and theoretical predictions of high N:P allocation strategies during resource competition[22].

The introduction of stoichiometric diversity allows the model to achieve the observed ΣN/ΣP values across a wide range of plausible denitrification and Fe-limitation scenarios (Fig. 1d). This increase is driven by the high N:P requirements of oligotrophic phytoplankton, which exacerbates N limitation in the subtropical gyres. A larger excess of PO_4^{3-} then remains to fuel N_2 fixation, expanding the niche of diazotrophs beyond that created by subsurface denitrification. This allows a balanced N budget to be achieved at higher values of ΣN/ΣP. Under weak Fe limitation and low denitrification rates, ΣN/ΣP actually exceeds the average N:P of marine plankton (Fig. 1d)—a condition that cannot be attained in previous box-model depictions. The ocean's ratio of nutrient reservoirs thus depends not only on the mean N:P of its plankton, but also on their stoichiometric diversity.

To investigate the sensitivity of ΣN/ΣP to different levels of stoichiometric diversity, we varied the N:P quota of diatom and non-diatom endmember communities, while maintaining a constant average export ratio of 16N:1P. As plankton stoichiometry becomes more diverse, global ΣN/ΣP rises steadily (Fig. 2), reflecting the expansion of the diazotrophic niche caused by the high N:P requirements of plankton cohabiting subtropical regions (increased $R_{O,ST}$). However, for every increase in $R_{O,ST}$, the increase in ΣN/ΣP is only 40% as large, much less than would be required for diazotrophs to keep ΣN/ΣP in line with the needs of their local competitors. This implies that the ecological niche of diazotrophs is determined not only by local competition with high N:P plankton, but also by remote diatom-dominated communities with a lower N:P quota. These communities largely occupy different ocean biomes, so their stoichiometric signatures must be communicated over long distances by ocean circulation.

To illustrate the role of circulation in controlling ΣN/ΣP, we employ a three-box model[12] (Fig. 3a) in which the circulations that transport

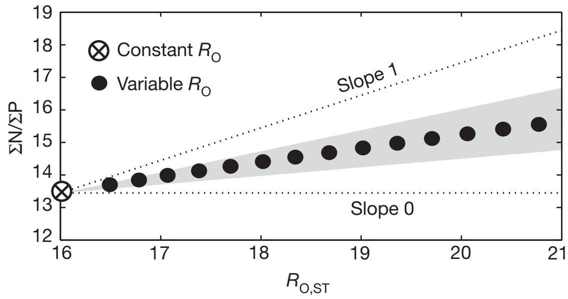


Figure 2 | Response of ΣN/ΣP to the degree of stoichiometric diversity, $R_{O,ST}$. Each simulation has 150 Tg N yr^{-1} denitrification and no Fe limitation. If the 'set point' of the nutrient thermostat were determined through local competition only, ΣN/ΣP would rise with $R_{O,ST}$ along a line of slope 1, yet a much weaker response is observed in our model. Grey shading represents an estimate of error for the slope of this line, derived using different estimates of diatom abundance as the basis for R_O (see Methods and Supplementary Notes).

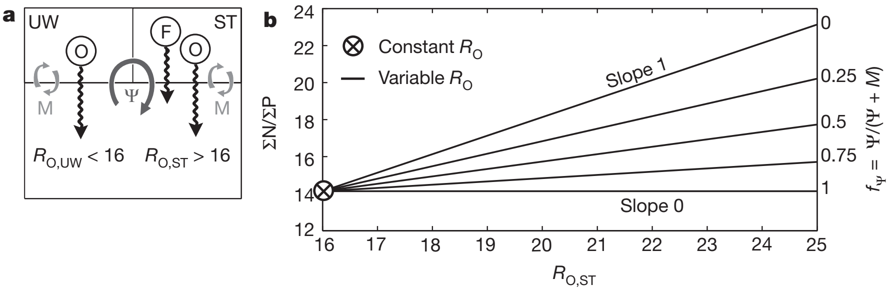


Figure 3 | Role of ocean circulation illustrated in a three-box model. **a**, Structure of the model, with two surface regions (upwelling UW; subtropics ST) and two circulation pathways (vertical exchange M; overturning Ψ). Diazotrophs are restricted to ST, and other plankton have diverse stoichiometry, which we vary by raising $R_{O,ST}$ and reducing $R_{O,UW}$ to maintain a mean of 16. **b**, The response of ΣN/ΣP to stoichiometric diversity depends on f_Ψ, the fraction of subtropical source waters that first pass through UW.

nutrients between the surface and deep ocean, and between diatom-dominated upwelling regions (UW) and downwelling subtropical gyres (ST), can be abstracted and manipulated (see Methods). When surface nutrients are supplied only from vertical exchange (M) with the deep ocean, the niche of diazotrophs is governed by their local competitors only, so every change in $R_{O,ST}$ produces an equal change in ΣN/ΣP, even though the global mean R_O is anchored at 16 (Fig. 3b, $f_\Psi = 0$). However, when the nutrients are predominantly supplied through shallow overturning and near-surface lateral circulations (Ψ), the signature of low N:P uptake—a higher residual NO_3^-:PO_4^{3-} ratio in surface waters—is transported directly from diatom-dominated upwelling regions into the subtropics. This reduces the excess PO_4^{3-} available to diazotrophs, and must be compensated by lower NO_3^-: PO_4^{3-} in upwelling deep water to maintain a given N_2-fixation rate. The response of the ocean's ΣN/ΣP to increases in $R_{O,ST}$ is thus damped as lateral circulations strengthen relative to vertical exchange. In the extreme case where all subtropical nutrients pass first through surface communities with low N:P ratios (Fig. 3b, $f_\Psi = 1$), the ocean N reservoir is entirely independent of spatial variations in R_O, and ΣN/ΣP is regulated through the global-mean N:P of plankton, as originally hypothesized by Redfield[1].

In light of these box model results, the tendency of ΣN/ΣP in the GCM to track only about 40% of a change in $R_{O,ST}$ (Fig. 2) suggests that about half the waters reaching subtropical sites of N_2-fixation are first influenced by low N:P plankton communities outside the subtropics. The dependence of global ΣN/ΣP on the spatial pattern of plankton N:P ratios can be computed by introducing taxon-dependent deviations from a constant-Redfield N:P (Fig. 1c, Supplementary Fig. 4) one grid cell at a time (see Methods). Regions can be divided between those whose stoichiometric deviations tend to raise ΣN/ΣP compared to the Redfieldian case (reddish positive values, Fig. 4a), and those that tend to reduce it (bluish negative values, Fig. 4a). Large negative values are found in the northern and equatorial Pacific and, to a lesser extent, in the polar regions of the Southern Ocean. The transport of nutrients from these source regions into the subtropics, through surface Ekman currents and shallow overturning, creates a 'biogeochemical teleconnection' by which remote low N:P communities reduce the availability of PO_4^{3-} to fuel N_2 fixation. This counteracts the expanded niche of diazotrophs produced by local competition with high-N:P subtropical communities, offsetting roughly half of the upwards pressure on ΣN/ΣP (Fig. 4b). Ocean circulation thus plays a critical role in the nutrient thermostat, reducing the bias of ΣN/ΣP towards the stoichiometry of subtropical communities, and holding the nutrient content of sea water closer to the global average requirements of phytoplankton.

The ratio of oceanic N and P reservoirs is a simple but powerful observational constraint on the dynamics of the marine N cycle. It appears to be fundamentally incompatible with a universal Redfield ratio of plankton biomass, lending global support to the large-scale association between biogeography and plankton nutrient metabolism inferred from Southern Ocean nutrient data[4].

The modern ocean's ΣN/ΣP also places new bounds on global denitrification rates and the limitations to diazotroph growth. Denitrification rates at the upper end of the estimated range (>250 Tg N yr^{-1}) are unable to yield the observed ΣN/ΣP ratio, even with a high degree of stoichiometric diversity (Fig. 1d). At the same time, the net effect of stoichiometric diversity among plankton taxa is to expand the ecological niche of marine diazotrophs. In geochemical estimates of N_2 fixation based on surface nutrients[23,24], this would translate into higher diagnosed N_2-fixation rates in the subtropics, not lower rates[24]. Thus, stoichiometric diversity helps to close the long-standing gap between estimates of N sources and sinks[25].

From a mechanistic perspective, the expanded niche for diazotrophs yields a higher ΣN/ΣP for a given denitrification rate, but this is still insufficient to achieve observed ΣN/ΣP when diazotrophs are strongly limited by airborne Fe (Fig. 1d). At most, the overall growth-rate handicap of diazotrophs can approach 50%, and then only if denitrification rates are at the lower end of the estimated range. If the intrinsic cost of diazotrophy were greater than the conservative value we use ($\mu_F/\mu_O = 0.95$), or if diazotroph growth is also slowed by

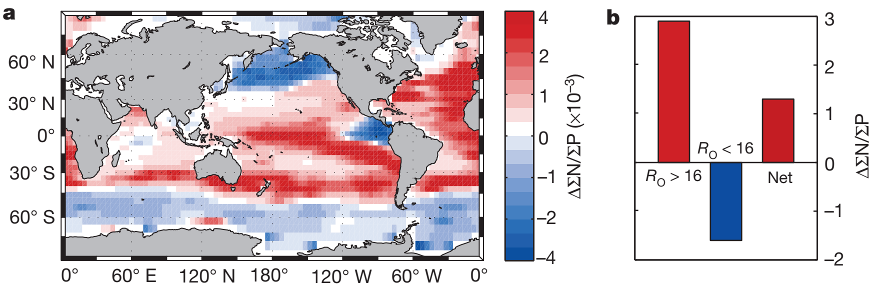


Figure 4 | Influence of individual surface regions on ΣN/ΣP. **a**, Values of ΔΣN/ΣP represent the change in steady-state ΣN/ΣP (from the Redfield case) prompted by introducing the grid cell's deviation of R_O from 16 (Supplementary Fig. 4), while holding $R_O = 16$ elsewhere (see Methods). **b**, Integral of ΔΣN/ΣP over regions of high R_O (>16) and low R_O (<16), and over the entire global domain ('Net'). High- and low-R_O regions exert opposite pressures on the ocean N reservoir, with a net increase in ΣN/ΣP over the Redfield case.

other factors not included here[6], then the observed ΣN/ΣP would require even weaker Fe limitation (see Supplementary Notes). These findings imply a secondary role for atmospheric Fe deposition in controlling rates of N_2 fixation in the modern ocean.

Temporal changes in the mean N:P of plankton have been hypothesized to drive long-term trends in ocean fertility and carbon storage, by shifting the set point of its nutrient thermostat[26,27]. Our model shows that this set point is also controlled by the biogeography of distinct plankton taxa, and ocean circulation patterns that transport nutrients between biomes—two factors known to vary with climatic conditions. Future stratification of the upper ocean and the expansion of oligotrophic biomes expected under a warming climate[28] could reshape the ecological niche of diazotrophs, and initiate a long-term perturbation in the ocean's nutrient thermostat.

METHODS SUMMARY

We used an observationally constrained ocean GCM (ref. 29) with horizontal resolution of 4° × 4° and 24 vertical layers, and simulated tracer transport using the transport matrix method[30]. A simple ecosystem model was adopted from refs 5 and 12, but modified for a three-dimensional global domain. The model includes four prognostic variables: NO_3^-, PO_4^{3-}, 'general' phytoplankton, and diazotrophic phytoplankton. Plankton growth rates vary as a function of temperature, light, and nutrient concentrations, and the parameters governing these relationships were tuned to optimize surface nutrient distributions (Supplementary Fig. 5). Diazotroph growth rates are scaled by atmospheric dust deposition to represent a heightened requirement for Fe in maintaining the nitrogenase enzyme. The strength of this scaling is varied to simulate differing degrees of Fe limitation. The spatial pattern of denitrification is governed by the degradation of organic matter in benthic grid cells and those where observed oxygen concentrations fall below 5 μM. The fluxes are then scaled in order to vary the global rate of N loss in a controlled manner between model scenarios. Each simulation was initialized with observed nutrient distributions and integrated for at least 10,000 years, until the N budget was balanced to within 0.1 Tg N yr^{-1}. Stoichiometric diversity among phytoplankton was parameterized using observed distributions of silicic acid to estimate the contribution of diatoms to nutrient export, and assuming different biomass N:P ratios for diatoms and other non-diazotrophic taxa. For each simulation, the value of $R_{O,ST}$ was calculated as the mean N:P ratio of plankton communities cohabiting surface grid boxes with diazotrophs at steady state.

Full Methods and any associated references are available in the online version of the paper.

Received 5 November 2011; accepted 27 June 2012.

1. Redfield, A. C. The biological control of chemical factors in the environment. *Am. Sci.* **46,** 205–221 (1958).
2. Redfield, A. C., Ketchum, B. H. & Richards, F. A. in *The Sea* Vol. 2 (ed. Hill, M. N.) 26–77 (Interscience, 1963).
3. Deutsch, C. & Weber, T. Nutrient ratios as a tracer and driver of ocean biogeochemistry. *Annu. Rev. Mar. Sci.* **4,** 113–141 (2012).
4. Weber, T. S. & Deutsch, C. Ocean nutrient ratios governed by plankton biogeography. *Nature* **467,** 550–554 (2010).
5. Tyrrell, T. The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature* **400,** 525–531 (1999).
6. Karl, D. *et al.* Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58,** 47–98 (2002).
7. Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B. & Carpenter, E. J. Trichodesmium, a globally significant marine cyanobacterium. *Science* **276,** 1221–1229 (1997).
8. Codispoti, L. A. in *Productivity of the Ocean: Past and Present* (eds Berger, W. H., Smetacek, V. S. & Wefer, G.) 377–394 (John Wiley and Sons, 1989).
9. Ward, B. B. *et al.* Denitrification as the dominant nitrogen loss process in the Arabian Sea. *Nature* **461,** 78–81 (2009).
10. Lam, P. & Kuypers, M. M. M. Microbial nitrogen cycling processes in oxygen minimum zones. *Annu. Rev. Mar. Sci.* **3,** 317–345 (2011).
11. Lenton, T. M. & Watson, A. J. Redfield revisited. 1. Regulation of nitrate, phosphate, and oxygen in the ocean. *Glob. Biogeochem. Cycles* **14,** 225–248 (2000).
12. Lenton, T. M. & Klausmeier, C. A. Biotic stoichiometric controls on the deep ocean N: P ratio. *Biogeosciences* **4,** 353–367 (2007).
13. Quigg, A. *et al.* The evolutionary inheritance of elemental stoichiometry in marine phytoplankton. *Nature* **425,** 291–294 (2003).
14. Green, S. E. & Sambrotto, R. N. Plankton community structure and export of C, N, P and Si in the Antarctic Circumpolar Current. *Deep Sea Res. II* **53,** 620–643 (2006).
15. Arrigo, K. R. *et al.* Phytoplankton community structure and the drawdown of nutrients and CO_2 in the Southern Ocean. *Science* **283,** 365–367 (1999).
16. Berman-Frank, I., Cullen, J. T., Shaked, Y., Sherrell, R. M. & Falkowski, P. G. Iron availability, cellular iron quotas, and nitrogen fixation in *Trichodesmium*. *Limnol. Oceanogr.* **46,** 1249–1260 (2001).
17. Kustka, A., Carpenter, E. J. & Sanudo-Wilhelmy, S. A. Iron and marine nitrogen fixation: progress and future directions. *Res. Microbiol.* **153,** 255–262 (2002).
18. DeVries, T., Deutsch, C., Primeau, F., Chang, B. & Devol, A. Global rates of water-column denitrification derived from nitrogen gas measurements. *Nature Geosci.* **5,** 547–550 (2012).
19. Heldal, M., Scanlan, D. J., Norland, S., Thingstad, F. & Mann, N. H. Elemental composition of single cells of various strains of marine *Prochlorococcus* and *Synechococcus* using X-ray microanalysis. *Limnol. Oceanogr.* **48,** 1732–1743 (2003).
20. Bertilsson, S., Berglund, O., Karl, D. M. & Chisholm, S. W. Elemental composition of marine *Prochlorococcus* and *Synechococcus*: implications for the ecological stoichiometry of the sea. *Limnol. Oceanogr.* **48,** 1721–1731 (2003).
21. Karl, D. M. *et al.* Ecological nitrogen-to-phosphorus stoichiometry at station ALOHA. *Deep Sea Res. II* **48,** 1529–1566 (2001).
22. Klausmeier, C. A., Litchman, E., Daufresne, T. & Levin, S. A. Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton. *Nature* **429,** 171–174 (2004).
23. Deutsch, C., Sarmiento, J. L., Sigman, D. M., Gruber, N. & Dunne, J. P. Spatial coupling of nitrogen inputs and losses in the ocean. *Nature* **445,** 163–167 (2007).
24. Mills, M. M. & Arrigo, K. R. Magnitude of oceanic nitrogen fixation influenced by the nutrient uptake ratio of phytoplankton. *Nature Geosci.* **3,** 412–416 (2010).
25. Codispoti, L. A. Biogeochemical cycles—is the ocean losing nitrate? *Nature* **376,** 724 (1995).
26. Falkowski, P. G. Rationalizing elemental ratios in unicellular algae. *J. Phycol.* **36,** 3–6 (2000).
27. Broecker, W. S. & Henderson, G. M. The sequence of events surrounding Termination II and their implications for the cause of glacial-interglacial CO_2 changes. *Paleoceanography* **13,** 352–364 (1998).
28. Polovina, J. J., Howell, E. A. & Abecassis, M. Ocean's least productive waters are expanding. *Geophys. Res. Lett.* **35,** L03618 (2008).
29. DeVries, T. & Primeau, F. Dynamically and observationally constrained estimates of water-mass distributions and ages in the global ocean. *J. Phys. Oceanogr.* **41,** 2381–2401 (2011).
30. Khatiwala, S. A computational framework for simulation of biogeochemical tracers in the ocean. *Glob. Biogeochem. Cycles* **21,** doi:10.1029/2007GB002923 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. DeVries for providing the ocean circulation model. This work was funded by a NASA Earth Systems Science Fellowship (T.W.) and a grant from the Gordon and Betty Moore Foundation (C.D.).

Author Contributions T.W. developed the model and performed simulations and analyses. Both authors designed the study and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.W. (tweber@atmos.ucla.edu).

METHODS

Circulation model. We use an observationally constrained ocean GCM which optimizes circulation to fit the linearized momentum equations, and observed distributions of temperature and salinity, and ^{14}C (ref. 29). It has horizontal resolution of $4^\circ \times 4^\circ$, and 24 vertical layers including two in the top 75 m. Annual-mean flow fields are extracted as a matrix, $\boldsymbol{A}$, facilitating tracer simulations using the transport matrix method[30].

Ecosystem model. We adopt a simple ecosystem model (similar to refs 5 and 12) that includes four prognostic variables, representing the concentrations (in μM) of NO_3^- (N), PO_4^{3-} (P), 'general' phytoplankton (O), and diazotrophic phytoplankton (F). Both phytoplankton types are simulated as organic phosphorous pools. The variables are governed by:

$$\frac{\mathrm{d}O}{\mathrm{d}t} = \mu_{\mathrm{O}} \min\left(\frac{P}{P+K_{\mathrm{P}}}, \frac{N}{N+K_{\mathrm{N}}}\right) O - MO \qquad (1)$$

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \mu_{\mathrm{F}} \frac{P}{P+K_{\mathrm{P}}} F - MF \qquad (2)$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = AP - (J_{\mathrm{O,UP}} + J_{\mathrm{F,UP}}) + Q_{\mathrm{rem}}(M(O + F)) \qquad (3)$$

$$\frac{\mathrm{d}N}{\mathrm{d}t} = AN - R_{\mathrm{O}} J_{\mathrm{O,UP}} + Q_{\mathrm{rem}}(M(R_{\mathrm{O}}O + R_{\mathrm{F}}F)) - D \qquad (4)$$

The parameters of the ecosystem model are discussed below and their numeric values listed in Supplementary Table 1.

Growth and mortality. The first terms on the right hand side of equations (1) and (2) represent plankton growth, as a function of environmental factors. The maximum growth rate of O (μ_{O}) is given by:

$$\mu_{\mathrm{O}}(T, I) = \mu_{\mathrm{opt}} \exp(k(T - T'))(1 - \exp(-I/K_I)) \qquad (5)$$

Here, μ_{opt} is the growth rate under optimal conditions, and k, T', K_{I}, K_{P} and K_{N} control the sensitivity of growth to temperature (T), light (I) and nutrient concentrations. Sensitivity parameters are tuned to reproduce observed surface nutrient distributions[31] in the model, ensuring a realistic pattern of biological nutrient drawdown (Supplementary Fig. 5). We account for the competitive handicap of diazotrophs by reducing their maximum growth rate (μ_{F}) relative to general plankton. It is scaled by a constant factor (δ_{F}), representing an intrinsic energetic expenditure on nitrogenase activity, and by a Fe-limitation parameter to represent the heightened requirement by diazotrophs for Fe:

$$\mu_{\mathrm{F}}(\mathrm{T,I,Fe}) = \mu_{\mathrm{O}} \delta_{\mathrm{F}} \frac{J_{\mathrm{Fe}}}{J_{\mathrm{Fe}} + K_{\mathrm{Fe}}} \qquad (6)$$

J_{Fe} is the simulated distribution of atmospheric Fe deposition onto the surface ocean[32], and the strength of the Fe-limitation is varied through K_{Fe} (Supplementary Fig. 1). In equations (1) and (2), M represents phytoplankton mortality, and includes a quadratic term that scales with total biomass ($M = m_1 + m_2B$, where $B = O + F$, and m_1 and m_2 are rate constants) and can be thought of as representing grazing by zooplankton, which are not explicitly simulated.

Nutrient cycling. Nutrients are transported by the circulation operator (matrix $\boldsymbol{A}$), and are assimilated into biological pools in the top 75 m through plankton growth. In equations (3) and (4), $J_{\mathrm{O,UP}}$ and $J_{\mathrm{F,UP}}$ are the same as the first terms on the right hand side of equations (1) and (2) respectively. R_{O} is the biomass N:P ratio of general phytoplankton, and R_{F} is the amount of N fixed per unit P uptake by diazotrophs. Following phytoplankton mortality, the recycling and remineralization of organic matter is simulated using the operator Q_{rem}. The majority is recycled in the surface ocean, and restored to local inorganic pools. A small fraction (ϕ_{e}) is exported from the surface layers as organic particles, and remineralized over depth following a power-law relation[33].

N budget. Newly fixed N is assumed to derive from an abundant dissolved N_2 pool that is not simulated explicitly. In equation (4), D represents the sum of water-column and sediment denitrification, which are simulated as sinks of NO_3^-. Water-column denitrification is proportional to the remineralization rate of organic matter in grid cells with climatological oxygen concentrations below a critical threshold $[O_2]_{\mathrm{crit}}$. Sediment denitrification is determined by the flux of organic matter to seafloor grid cells[34]. Because global denitrification rates are one of the primary determinants of the steady-state N inventory, but are not well constrained observationally, D is scaled to maintain a specified global rate, and a constant partition among water column and sediments, thus controlling for these factors between simulations. Simulated distributions of N sources and sinks are shown in Supplementary Fig. 2.

Plankton N:P ratios. We assume that stoichiometric variability occurs primarily at the taxonomic level, and use the empirical relation derived by ref. 4:

$$R_{\mathrm{O}} = R_{\mathrm{O,diat}} \phi_{\mathrm{diat}} + R_{\mathrm{O,other}}(1 - \phi_{\mathrm{diat}}) \qquad (7)$$

Here, ϕ_{diat} is the fractional contribution of diatoms to nutrient export, and $R_{\mathrm{O,diat}}$ and $R_{\mathrm{O,other}}$ are the biomass N:P ratios of diatoms and other non-diazotropic phytoplankton respectively. Rather than simulate different taxonomic groups explicitly, we use a prior estimate of ϕ_{diat} combined with equation (7) to apply a spatially varying pattern of R_{O} to the single class of non-diazotrophic plankton. Three different approaches are considered for estimating ϕ_{diat} (Supplementary Fig. 3). Method 1 assumes that the relative abundance of diatoms scales with the observed surface concentration of $Si(OH)_4$ (ref. 35):

$$\phi_{\mathrm{diat}} = \frac{[\mathrm{Si}]}{[\mathrm{Si}] + K_{\mathrm{Si}}} \qquad (8)$$

K_{Si} is tuned to accommodate the observational constraint that diatoms contribute 40–50% of global export production[36]. Methods 2 and 3 compute ϕ_{diat} from the relative export fluxes of N and Si and an estimate of the Si:N ratios in diatom biomass[37]. Method 2 is based on observations only[38], diagnosing export fluxes from the vertical gradients of Si and N between the thermocline and surface, whereas Method 3 diagnoses the fluxes in an ocean GCM[36]. We note that the two diagnostic methods are less appropriate in regions where N_2 fixation confounds the diagnosis of N export. We used the simplest approach (Method 1), which agrees most closely with satellite-derived estimates of diatom biogeography[39]. Methods 2 and 3 are used to derive an estimate of uncertainty associated with the community-composition parameterization.

In our initial simulations (Fig. 1), $R_{\mathrm{O,diat}}$ and $R_{\mathrm{O,other}}$ are held close to the values diagnosed in ref. 4, with the added constraint that the mean N:P export ratio by non-fixing plankton is equal to the Redfield ratio of 16:1 (Supplementary Fig. 4). In later simulations (Fig. 2), we vary $R_{\mathrm{O,diat}}$ and $R_{\mathrm{O,other}}$ to produce different degrees of stoichiometric diversity, but again ensure the global-mean constraint is satisfied. This allows us to identify changes in ΣN/ΣP that are caused only by changes in the spatial pattern of R_{O}, and not its global-mean value. The ratio of N_2 fixation to P uptake by diazotrophs (R_{F}) is assumed to be constant, but as in previous studies[12], our results are not sensitive to the value of this parameter.

Sensitivity testing. We rigorously tested the sensitivity of our results to parameters and assumptions of the ecosystem model. See Supplementary Notes, Supplementary Figs 7–10 and Supplementary Table 2.

Three-box model. We use a three-box model of the ocean to assist our interpretation of the ocean GCM results. The geometry and nutrient fluxes of the model are shown in Supplementary Fig. 6. Its ecosystem and biogeochemistry components are held as close to the ocean GCM version as possible for ease of comparison, with the following simplifications:

(1) Phytoplankton growth rates are set to the optimal value in ST, and reduced by a factor of 0.5 in the UW, so that residual nutrients remain in the surface as observed.

(2) The competitive handicap of diazotrophs is determined only by δ_{F} in ST (no Fe limitation), and μ_{F} is set to zero in UW.

(3) Denitrification is distributed between the surface and deep ocean as in ref. 5.

Regional control of ΣN/ΣP. For each surface grid cell (with x,y coordinates i,j), a simulation was conducted in which the local R_{O} was set to its stoichiometrically diverse value, as computed from ϕ_{diat} (Supplementary Fig. 4). In all other surface regions, R_{O} was held equal to the Redfield ratio. The difference between the steady-state ΣN/ΣP in this simulation, and that in a uniform Redfieldian case, was then computed:

$$\Delta \Sigma N/\Sigma P\big|_{\substack{x=i\\y=j}} = \frac{\Sigma N}{\Sigma P}\left(\begin{cases} R_{\mathrm{O}} = f(\phi_{\mathrm{diat}}), x = i, y = j \\ R_{\mathrm{O}} = 16, x \neq i, y \neq j \end{cases}\right) - \frac{\Sigma N}{\Sigma P}(R_{\mathrm{O}} = 16) \qquad (9)$$

This value is taken as a measure of the sensitivity of ΣN/ΣP to the uptake stoichiometry of plankton communities in the perturbed surface region. The efficiency of these computations was enhanced using a quasi-steady-state assumption for the 'fast' biological variables O and F, which reduces the model to a two-equation system that can be solved directly for steady state using Newton's method[40]. Sensitivity testing demonstrated that the solutions derived from the quasi-steady-state-assumption approach and full four-equation model were almost indistinguishable.

31. Garcia, H. E., Locarni, R. A., Boyer, T. P. & Antonov, J. I. *World Ocean Atlas 2005* Vol. 4 *Nutrients (phosphate, nitrate, silicate)* (US Government Printing Office, 2006).

32. Mahowald, N. M. *et al.* Change in atmospheric mineral aerosols in response to climate: last glacial period, preindustrial, modern, and doubled carbon dioxide climates. *J. Geophys. Res.* **111,** doi:10.1029/2005JD006653 (2006).

33. Martin, J. H., Gordon, R. M., Fitzwater, S. & Broenkow, W. W. VERTEX: phytoplankton/iron studies in the Gulf of Alaska. *Deep-Sea Res.* **36,** 649–680 (1989).
34. Middelburg, J. J., Soetaert, K., Herman, P. M. J. & Heip, C. H. R. Denitrification in marine sediments: a model study. *Glob. Biogeochem. Cycles* **10,** 661–673 (1996).
35. Egge, J. K. & Aksnes, D. L. Silicate as regulating nutrient in phytoplankton competition. *Mar. Ecol. Prog. Ser.* **83,** 281–289 (1992).
36. Jin, X., Gruber, N., Dunne, J. P., Sarmiento, J. L. & Armstrong, R. A. Diagnosing the contribution of phytoplankton functional groups to the production and export of particulate organic carbon, $CaCO_3$, and opal from global nutrient and alkalinity distributions. *Glob. Biogeochem. Cycles* **20,** doi:10.1029/2005GB002532 (2006).
37. Brzezinski, M. A. *et al.* A switch from $Si(OH)_4$ to NO_3-depletion in the glacial Southern Ocean. *Geophys. Res. Lett.* **29,** 1564 (2002).
38. Sarmiento, J. L. & Gruber, N. *Ocean Biogeochemical Dynamics* (Princeton University Press, 2006).
39. Alvain, S., Moulin, C., Dandonneau, Y. & Loisel, H. Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: a satellite view. *Glob. Biogeochem. Cycles* **22,** doi:10.1029/2007GB003154 (2008).
40. Kwon, E. Y. & Primeau, F. Optimization and sensitivity study of a biogeochemistry ocean model using an implicit solver and *in situ* phosphate data. *Glob. Biogeochem. Cycles* **20,** doi:10.1029/2005GB002631 (2006).

doi:10.1038/nature11377

Afternoon rain more likely over drier soils

Christopher M. Taylor[1], Richard A. M. de Jeu[2], Françoise Guichard[3], Phil P. Harris[1] & Wouter A. Dorigo[4]

Land surface properties, such as vegetation cover and soil moisture, influence the partitioning of radiative energy between latent and sensible heat fluxes in daytime hours. During dry periods, soil-water deficit can limit evapotranspiration, leading to warmer and drier conditions in the lower atmosphere[1,2]. Soil moisture can influence the development of convective storms through such modifications of low-level atmospheric temperature and humidity[1,3], which in turn feeds back on soil moisture. Yet there is considerable uncertainty in how soil moisture affects convective storms across the world, owing to a lack of observational evidence and uncertainty in large-scale models[4]. Here we present a global-scale observational analysis of the coupling between soil moisture and precipitation. We show that across all six continents studied, afternoon rain falls preferentially over soils that are relatively dry compared to the surrounding area. The signal emerges most clearly in the observations over semi-arid regions, where surface fluxes are sensitive to soil moisture, and convective events are frequent. Mechanistically, our results are consistent with enhanced afternoon moist convection driven by increased sensible heat flux over drier soils, and/or mesoscale variability in soil moisture. We find no evidence in our analysis of a positive feedback—that is, a preference for rain over wetter soils—at the spatial scale (50–100 kilometres) studied. In contrast, we find that a positive feedback of soil moisture on simulated precipitation does dominate in six state-of-the-art global weather and climate models—a difference that may contribute to excessive simulated droughts in large-scale models.

Soil moisture influences precipitation across a range of scales in time and space[5]. In drought-affected continental regions, weak evapotranspiration leads to reduced atmospheric moisture content over a period of days, potentially suppressing subsequent precipitation[6]. When soil moisture anomalies are extensive, surface-induced perturbations to the atmospheric heat budget may modify synoptic-scale circulations[2], in turn affecting moisture advection from the oceans[7]. On smaller scales, the development of convective clouds and precipitation can be influenced by local surface fluxes over the course of the day[1,3]. Theoretical considerations[8,9] suggest that, in an undisturbed atmosphere, the likelihood and sign of a surface feedback will be determined by the atmospheric profiles of temperature and humidity. Thus, one might expect regional variations in the strength and sign of convective sensitivity to soil moisture[10,11]. Mesoscale variability in soil moisture can also influence the feedback through the development of daytime circulations[12], which provide additional convergence to trigger convection[13,14].

Several studies have examined the impact of the land surface on observed rainfall in different regions of the world. Analyses in Illinois[15] and West Africa[16] have indicated positive correlations between antecedent soil moisture and precipitation, consistent with a positive soil moisture feedback. A recent study[17] based on observationally constrained reanalysis data showed an increasing frequency of convective rainfall when evapotranspiration was higher across much of North America. On the other hand, examination of satellite cloud data has indicated locally enhanced afternoon precipitation frequency over surfaces with increased sensible heat fluxes, as a result of mesoscale circulations due either to soil moisture[18] or vegetation cover[19,20].

At the regional scale, climate models tend to agree on where feedbacks occur, these being constrained largely by where soil moisture limits evapotranspiration in the presence of convective activity[4]. But the spread in simulated feedback strength is large, highlighting both the uncertainty in surface flux sensitivity to soil moisture and the response of the planetary boundary layer and convection to surface fluxes[21,22]. Indeed, the feedback sign can change depending on model spatial resolution, with a strong influence of the convective parameterization likely to be responsible[23].

Until recently, there has been a lack of observations with which to evaluate feedbacks in large-scale models. We address that problem here, and focus on the least well understood aspect of the feedback loop between soil moisture and precipitation, namely, the response of daytime moist convection to soil moisture anomalies. In the past decade, global observational data sets of both surface soil moisture[24,25] and precipitation[26] have become available at a resolution of $0.25^\circ \times 0.25^\circ$, on daily and 3-hourly time steps respectively. We use these to analyse the location of afternoon rain events relative to the underlying antecedent soil moisture. In particular we examine whether rain is more likely over soils that are wetter or drier than the surrounding area.We then apply the same methodology to six global models used in reanalyses or climate projections.

We focus on the development of precipitation events during the afternoon, when the sensitivity of convection to land conditions is expected to be maximized. An event is defined at a $0.25^\circ \times 0.25^\circ$ pixel location ($L_{\max}$) with a maximum in afternoon rainfall, centred in a box measuring $1.25^\circ \times 1.25^\circ$ (see Methods Summary and Supplementary Fig. 3). Each $L_{\max}$ is paired with one or more pixels in the box where afternoon rainfall is at a minimum ($L_{\min}$). We compute the difference in pre-rain-event soil moisture, ΔS_e, between $L_{\max}$ and $L_{\min}$ having first subtracted a climatological mean soil moisture from both locations. We quantify the strength of the soil moisture effect on precipitation using a sample of events, and assess how unexpected the observed sample mean value of ΔS_e is, relative to a control sample, ΔS_c, from the same location pairs on non-event days. More precisely, we examine the difference in ΔS between the event and control samples, $\delta_e = \text{mean}(\Delta S_e) - \text{mean}(\Delta S_c)$, expressed as a percentile of typical δ values (see Methods Summary). Mountainous and coastal areas are excluded because of their effects on mesoscale precipitation, and we are unable to analyse the observations in tropical forest regions, owing to the limitations of soil moisture retrievals beneath dense vegetation.

The map in Fig. 1 shows regions of the world where afternoon precipitation is observed more frequently than expected over wet (blue) or dry (red) soils, based on analysis of δ_e at a scale of 5°. Globally, 28.9% of the grid cells analysed have percentile values, P, less than 10, as compared to an expected frequency (assuming no feedback) of 10%, and just 3.4% with $P > 90$. Clusters of low percentiles are found in semi-arid and arid regions, most notably North Africa, but also in Eastern Australia, Central Asia and Southern Africa. These clusters indicate a clear preference for afternoon rain over drier soils

[1] NERC Centre for Ecology and Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford OX10 8BB, UK. [2] Vrije Universiteit Amsterdam, Earth and Climate Cluster, Department of Earth Sciences, Faculty of Earth and Life Sciences, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands. [3] CNRM-GAME (CNRS and Météo-France), 42 avenue Coriolis, 31057 Toulouse Cedex, France. [4] Institute of Photogrammetry and Remote Sensing (I.P.F.), Vienna University of Technology, Gusshausstraße 27-29, 1040 Vienna, Austria.

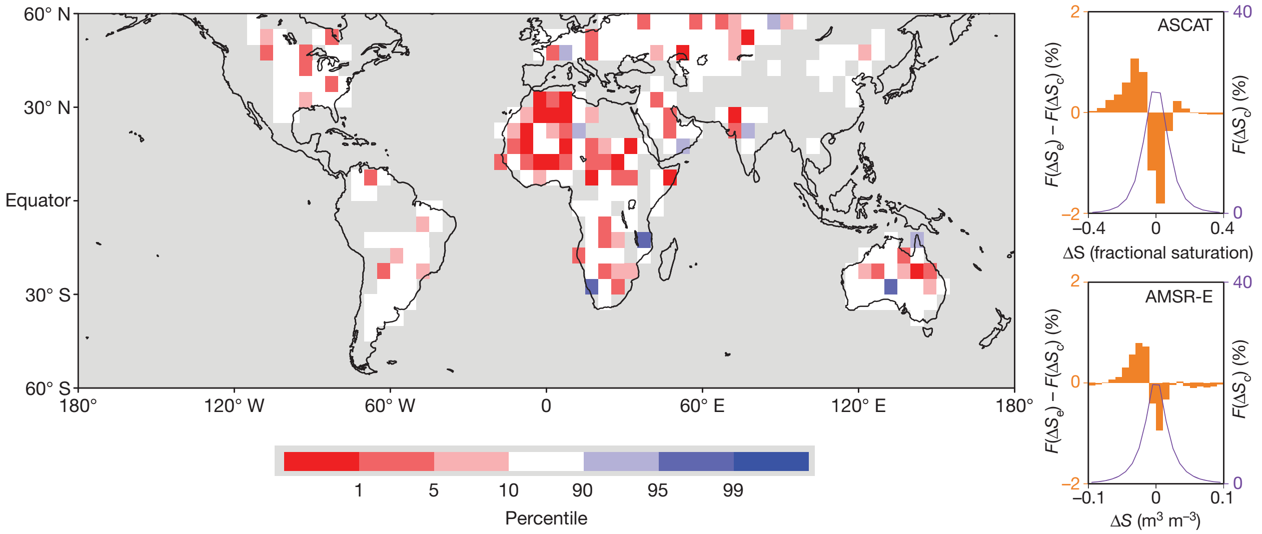


Figure 1 | Preference for afternoon precipitation over soil moisture anomalies. Percentiles of the observed variable $\delta_e = \text{mean}(\Delta S_e) - \text{mean}(\Delta S_c)$ for each 5° × 5° box under a null assumption that no feedback exists. Null sampling distributions of δ values were estimated for each box by re-sampling without replacement from the combined set of event and non-event ΔS values. Low (high) percentiles indicate where rainfall maxima occur over locally dry (wet) soil more frequently than expected. Grey denotes 5° × 5° cells containing fewer than 25 events. The map is based on a merging of two separate analyses using either ASCAT or AMSR-E soil moisture. For each 5° × 5° cell, the relative quality of the two data sets is tested independently to determine which product is used (Supplementary Figs 5, 6). Insets: frequency histograms $F(\Delta S_c)$ of soil moisture difference in the global control sample (purple), and the difference $F(\Delta S_e) - F(\Delta S_c)$ between the histograms of the global event and global control samples (orange shading). The total number of events (n_e) is 29,729 for ASCAT and 73,623 for AMSR-E. Note the different units for ΔS for ASCAT (fractional saturation) and AMSR-E ($m^3 m^{-3}$).

in those regions, consistent with a previous study over the Western Sahel[18]. This signal is also evident when computing δ_e from all events across the world (Fig. 1 insets). Further analysis (Supplementary Information and Supplementary Tables 3 and 4) demonstrates that this signal is statistically significant at the 99% level over all continents and in all climate zones, with the exception of tropical forests, where accurate soil moisture retrievals are unavailable. We repeated the analysis after degrading the spatial resolution from 0.25° to 1.0° . This produced only about one-tenth of the number of events identified in the 0.25° data, but a statistically robust preference for rain over drier soil was still found across the tropics, and in particular over parts of North Africa and Australia (Supplementary Fig. 10; Supplementary Tables 3, 4).

Using two alternative precipitation data sets, we found the same global preference for rain over drier soil, and similar regions contributing to that signal (Supplementary Fig. 8; Supplementary Tables 3, 4). Although all of the satellite-derived data sets are subject to errors at the event scale, analysing the data over many events should yield more accurate estimates of δ_e. Furthermore, our approach exploits an aspect of rainfall that is relatively well captured by satellite, that is, its spatial structure. Additional analysis (Supplementary Fig. 4) indicates a strong degree of mutual consistency in the spatial variability of soil moisture and rainfall in our independent data sets, providing further evidence to support our methodology.

We now consider whether the observed preference for rain over drier soil is consistent with land surface feedback. For a soil moisture feedback on precipitation, soil water deficit must limit evapotranspiration. This regime is found only in certain seasons and regions of the world[4], where water stress coincides with convective activity. Low percentiles in Fig. 1 occur in areas that are relatively dry, and originate from seasons with convective storms (Supplementary Fig. 9). Using data from across the globe, the sensitivity of δ_e to the areal-mean (1.25° × 1.25°) soil moisture is explored in Fig. 2a. The most negative values (rain over drier soil) are found for the driest mean conditions, and the signal loses significance at the 95% level above 0.20 $m^3 m^{-3}$. This behaviour is consistent with soil moisture feedback, as the sensitivity of sensible and latent heat fluxes to soil moisture increases as mean soil moisture decreases. Also, the use of surface soil moisture

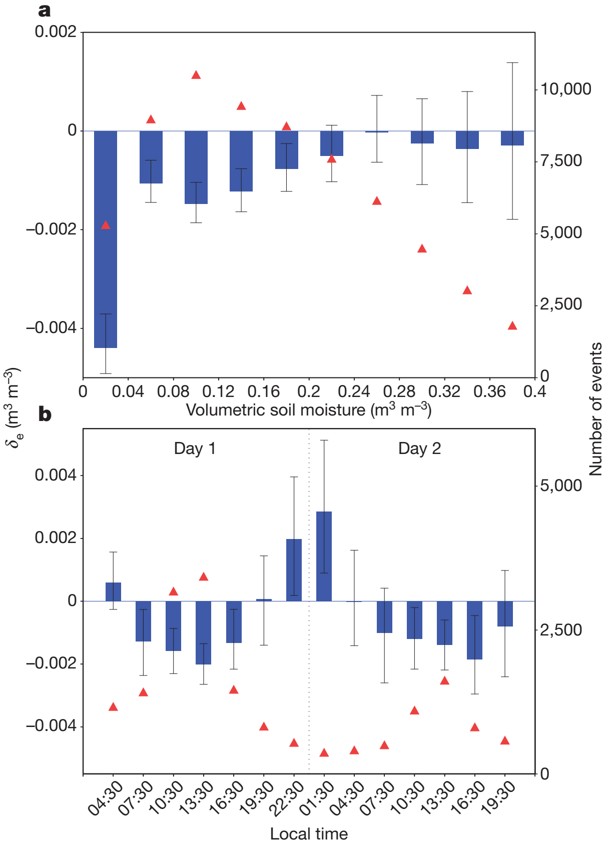


Figure 2 | Sensitivities of pre-rain-event soil moisture to mean soil moisture and time of day. Blue bars denote the anomalous pre-rain-event soil moisture difference, δ_e, averaged over every event globally, as a function of pre-event soil moisture averaged over 1.25° × 1.25° (**a**), and time of first precipitation (at least 1 mm over 3 h), following a soil moisture measurement at 1:30 on day 1 (**b**). Negative values of δ_e indicate a preference for precipitation over drier soil, and error bars show 90% confidence limits. Red triangles denote the number of events used for each δ_e average.

as a proxy for surface flux variability should be most effective for dry and sparsely vegetated surfaces.

A land feedback requires a strong diurnal sensitivity in the observed signal. We repeated our analysis, this time detecting the onset of precipitation at varying lag times after a soil moisture observation at 1:30 (all times are local time). The values of δ_e (Fig. 2b) exhibit a pronounced diurnal cycle, still evident 36 hours after the observation. The most negative values occur during daytime, in particular between 12:00 and 15:00. By contrast, between 21:00 and 3:00 the opposite signal emerges; that is, events are more likely to be found over wetter soils. The early afternoon minimum is consistent with a negative soil moisture feedback on convective initiation, when the effects of surface properties on the planetary boundary layer, convective instability and mesoscale flows are all maximized. Mechanisms to explain the reverse signal in the hours around midnight may be more subtle. The effects of thermals and daytime surface-induced flows are likely to be relatively short-lived after dusk. On the other hand, nocturnal humidity anomalies may persist for longer, depending on the spatial scale of the surface features and wind conditions. From detailed examination of individual events, it appears that, overnight, there is an increasing influence of pre-existing, fast-moving convective systems in our sample, particularly in the Sahel. Distinct mechanisms will be involved in the surface interaction with organized convective systems, which may favour a positive feedback[16].

Finally, we repeat our analysis using 3-hourly diagnostics from six global models, ranging in horizontal resolution from 0.5 to 2.0°. Our results (Fig. 3) indicate a strong preference for rain over wet soils for large parts of the world, in contrast to the observations. Only one model (Fig. 3e) produces more than the expected 10% of grid cells with $P < 10$, largely due to contributions at mid-latitudes. The cross-model signal favouring precipitation over wet soil, particularly across the tropics (Supplementary Table 3), demonstrates a fundamental failing in the ability of convective parameterizations to represent land feedbacks on daytime precipitation. This is likely to be linked to the oft-reported phase lag in the diurnal cycle of precipitation; that is, simulated rainfall tends to start several hours too early[27], and is possibly amplified by a lack of boundary-layer clouds in some models. This weakness has been related to the crude criteria used to trigger deep convection in large-scale models[28]. The onset of convective precipitation is overly sensitive to the daytime increase of moist convective instability, which is typically faster over wetter soils[3], favouring a positive feedback. Early initiation limits the effect of other daytime processes on triggering convection in the models. In contrast, our observational analysis points to the importance of dry boundary-layer dynamics for this phenomenon over land.

The observed preference for afternoon rain over locally drier soil on scales of 50–100 km is consistent with a number of regional studies based on remotely sensed data[18–20]. Our failure to find areas of positive

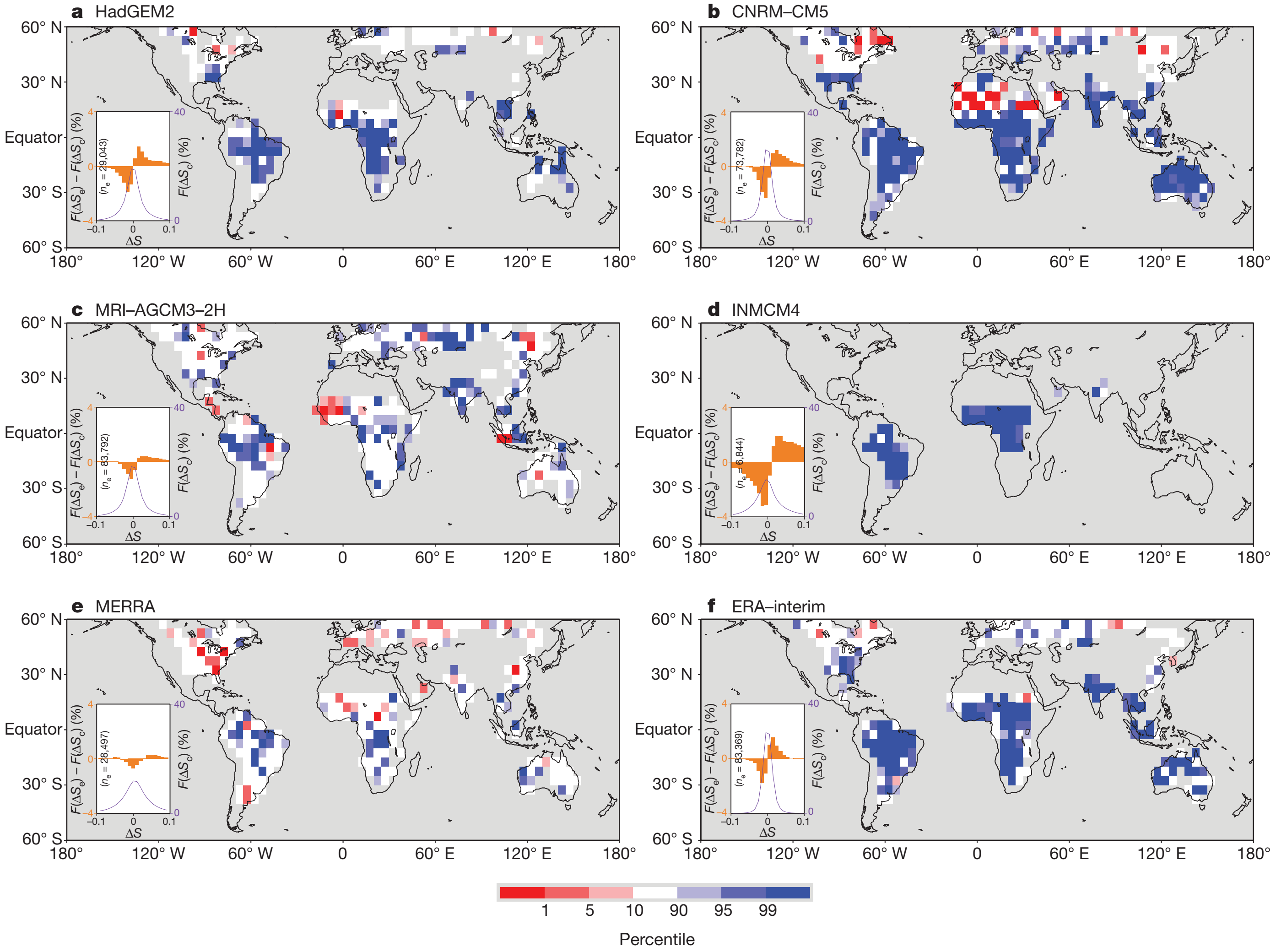


Figure 3 | Simulated preference for afternoon precipitation over soil moisture anomalies. As for Fig. 1 but using diagnostics from integrations by four climate models (**a**–**d**) and two atmospheric reanalysis models (**e**, **f**). Blue (red) shading indicates convective precipitation more likely over wetter (drier) soils. The models used are: **a**, HadGEM2; **b**, CNRM-CM5; **c**, MRI-AGCM3-2H; **d**, INMCM4; **e**, MERRA; and **f**, ERA-Interim. Inset as for Fig. 1, with ΔS in $m^3\,m^{-3}$. Further details of the models are in Supplementary Information, with maps of the number of events in each model in Supplementary Fig. 11.

feedback may indicate the importance of surface-induced mesoscale flows in triggering convection[18], although the coarse spatial resolution of our data sets prevents us from drawing firm conclusions on this issue. Equally, mixing processes in the growth stage of convective clouds before precipitation[23,29] may play an important role. Neither of these processes is captured in existing one-dimensional analyses[8]. Furthermore, our results raise questions about the ability of models reliant on convective parameterizations to represent these processes adequately. Although the coarser-resolution models analysed here (HadGEM2, CNRM-CM5 and INMCM4) cannot resolve mesoscale soil moisture structures, nor their potential impacts on convective triggering[18], all the models have a strong tendency towards rain over wetter soils, for which we find no observational support. Our study does not, however, imply that the soil moisture feedback is negative at temporal and spatial scales different from those analysed here. The multi-day accumulation of moisture in the lower atmosphere from a freely transpiring land surface may provide more favourable initial (dawn) conditions for daytime convection than the equivalent accumulation over a drought-affected region. Equally, the large-scale dynamical response to soil moisture may dominate in some regions. However, the erroneous sensitivity of convection schemes demonstrated here is likely to contribute to a tendency for large-scale models to 'lock-in' dry conditions, extending droughts unrealistically, and potentially exaggerating the role of soil moisture feedbacks in the climate system[30].

METHODS SUMMARY

Surface soil moisture retrievals are used between 60° S and 60° N from the Advanced Microwave Scanning Radiometer for EOS (AMSR-E; June 2002 to October 2011)[24], and the MetOP Advanced Scatterometer (ASCAT; 2007–11)[25]. They have typically one overpass per pixel per day at either 1:30 or 13:30 (AMSR-E), and 9:30 or 21:30 (ASCAT). Additional soil moisture quality control procedures are described in Supplementary Information. The CMORPH[26] 3-hourly precipitation data set is based on data from a combination of satellites.

Locations of afternoon events, $L_{\max}$, are defined within a box measuring 5×5 pixels by the maximum accumulated precipitation (12:00–21:00) that exceeds 3 mm. We exclude pixels with more than 1 mm rain in the preceding hours, and apply an additional filter to remove cases close to active precipitation when using soil moisture data for 13:30. These steps ensure that the soil moisture measurement precedes the rainfall. Locations where topographic height variability exceeds 300 m are excluded, along with regions containing water bodies or strong climatological soil moisture gradients.

The control sample, ΔS_c, is constructed from daily soil moisture differences between locations $L_{\max}$ and $L_{\min}$, using data for the same calendar month but from non-event years. This quantifies typical (non-event) soil moisture differences between the locations. Each value in samples ΔS_e and ΔS_c has an individual climatological mean ΔS subtracted, which is calculated from ΔS values in the same calendar month in non-event years. For the models, soil moisture and rainfall accumulations are available every 3 h (universal time). Because of the models' lower spatial resolution (0.5–2.0°), the event box is reduced to 3×3 pixels and the local time window between 6:00 and 8:59 adopted to calculate ΔS. Convective rain is accumulated in the subsequent 9 h, several hours in the day earlier, to account for diurnal phase bias in model precipitation.

Received 19 March; accepted 29 June 2012.
Published online 12 September 2012.

1. Betts, A. K. & Ball, J. H. FIFE surface climate and site-average dataset 1987–89. *J. Atmos. Sci.* **55,** 1091–1108 (1998).
2. Fischer, E. M. *et al.* Soil moisture-atmosphere interactions during the 2003 European summer heat wave. *J. Clim.* **20,** 5081–5099 (2007).
3. Eltahir, E. A. B. A soil moisture-rainfall feedback mechanism. 1. Theory and observations. *Wat. Resour. Res.* **34,** 765–776 (1998).
4. Koster, R. D. *et al.* Regions of strong coupling between soil moisture and precipitation. *Science* **305,** 1138–1140 (2004).
5. Goessling, H. F. & Reick, C. H. What do moisture recycling estimates tell us? Exploring the extreme case of non-evaporating continents. *Hydrol. Earth Syst. Sci.* **15,** 3217–3235 (2011).
6. van der Ent, R. J., Savenije, H. H. G., Schaefli, B. & Steele-Dunne, S. C. Origin and fate of atmospheric moisture over continents. *Wat. Resour. Res.* **46,** W09525 (2010).
7. Webster, P. J. Mechanisms of monsoon low-frequency variability - surface hydrological effects. *J. Atmos. Sci.* **40,** 2110–2124 (1983).
8. Findell, K. L. & Eltahir, E. A. B. Atmospheric controls on soil moisture-boundary layer interactions. Part I: framework development. *J. Hydrometeorol.* **4,** 552–569 (2003).
9. Ek, M. B. & Holtslag, A. A. M. Influence of soil moisture on boundary layer cloud development. *J. Hydrometeorol.* **5,** 86–99 (2004).
10. Findell, K. L. & Eltahir, E. A. B. Atmospheric controls on soil moisture-boundary layer interactions. Part II: feedbacks within the continental United States. *J. Hydrometeorol.* **4,** 570–583 (2003).
11. Ferguson, C. R. & Wood, E. F. Observed land–atmosphere coupling from satellite remote sensing and reanalysis. *J. Hydrometeorol.* **12,** 1221–1254 (2011).
12. Ookouchi, Y., Segal, M., Kessler, R. C. & Pielke, R. A. Evaluation of soil moisture effects on the generation and modification of mesoscale circulations. *Mon. Weath. Rev.* **112,** 2281–2292 (1984).
13. Cheng, W. Y. Y. & Cotton, W. R. Sensitivity of a cloud-resolving simulation of the genesis of a mesoscale convective system to horizontal heterogeneities in soil moisture initialization. *J. Hydrometeorol.* **5,** 934–958 (2004).
14. Anthes, R. A. Enhancement of convective precipitation by mesoscale variations in vegetative covering in semi-arid regions. *J. Clim. Appl. Meteorol.* **23,** 541–554 (1984).
15. Findell, K. L. & Eltahir, E. A. B. An analysis of the soil moisture-rainfall feedback, based on direct observations from Illinois. *Wat. Resour. Res.* **33,** 725–735 (1997).
16. Taylor, C. M. & Lebel, T. Observational evidence of persistent convective-scale rainfall patterns. *Mon. Weath. Rev.* **126,** 1597–1607 (1998).
17. Findell, K. L., Gentine, P., Lintner, B. R. & Kerr, C. Probability of afternoon precipitation in eastern United States and Mexico enhanced by high evaporation. *Nature Geosci.* **4,** 434–439 (2011).
18. Taylor, C. M. *et al.* Frequency of Sahelian storm initiation enhanced over mesoscale soil-moisture patterns. *Nature Geosci.* **4,** 430–433 (2011).
19. Wang, J. F. *et al.* Impact of deforestation in the Amazon basin on cloud climatology. *Proc. Natl Acad. Sci. USA* **106,** 3670–3674 (2009).
20. Carleton, A. M. *et al.* Synoptic circulation and land surface influences on convection in the Midwest US "corn belt" during the summers of 1999 and 2000. Part II: role of vegetation boundaries. *J. Clim.* **21,** 3617–3641 (2008).
21. Santanello, J. A., Peters-Lidard, C. D. & Kumar, S. V. Diagnosing the sensitivity of local land–atmosphere coupling via the soil moisture–boundary layer interaction. *J. Hydrometeorol.* **12,** 766–786 (2011).
22. Guo, Z. C. *et al.* GLACE: The Global Land-Atmosphere Coupling Experiment. Part II: analysis. *J. Hydrometeorol.* **7,** 611–625 (2006).
23. Hohenegger, C., Brockhaus, P., Bretherton, C. S. & Schar, C. The soil moisture-precipitation feedback in simulations with explicit and parameterized convection. *J. Clim.* **22,** 5003–5020 (2009).
24. Owe, M., de Jeu, R. & Holmes, T. Multisensor historical climatology of satellite-derived global land surface moisture. *J. Geophys. Res.* **113,** F01002 (2008).
25. Bartalis, Z. *et al.* Initial soil moisture retrievals from the METOP-A Advanced Scatterometer (ASCAT). *Geophys. Res. Lett.* **34,** L20401 (2007).
26. Joyce, R. J., Janowiak, J. E., Arkin, P. A. & Xie, P. CMORPH: a method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeorol.* **5,** 487–503 (2004).
27. Dai, A. Precipitation characteristics in eighteen coupled climate models. *J. Clim.* **19,** 4605–4630 (2006).
28. Guichard, F. *et al.* Modelling the diurnal cycle of deep precipitating convection over land with cloud-resolving models and single-column models. *Q. J. R. Meteorol. Soc.* **130,** 3139–3172 (2004).
29. Zhang, Y. & Klein, S. A. Mechanisms affecting the transition from shallow to deep convection over land: inferences from observations of the diurnal cycle collected at the ARM southern Great Plains site. *J. Atmos. Sci.* **67,** 2943–2959 (2010).
30. McCrary, R. R. & Randall, D. A. Great Plains drought in simulations of the twentieth century. *J. Clim.* **23,** 2178–2196 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was partly funded by the European Union (FP6) WATCH Integrated Project (contract 036946), the UK National Centre for Earth Observation and the European Space Agency STSE Water Cycle Multi-mission Observation Strategy (WACMOS) project (ESRIN/contract number 22086/08/I-EC. We thank A. Beljaars, S. Seneviratne and D. Parker for discussions on this topic. We also thank the CMORPH, TRMM, PERSIANN and GPROF teams for the provision of their precipitation data, the World Climate Research Programme's Working Group on Coupled Modelling, and the centres who provided modelling data in Figure 3.

Author Contributions C.M.T. and R.A.M.d.J. conceived the study, C.M.T. performed the analysis and wrote the paper, R.A.M.d.J. and W.A.D. provided expertise on soil moisture data sets, F.G. interpreted the convective responses in models and observations, and P.P.H. devised statistical tests. All authors discussed the results and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.M.T. (cmt@ceh.ac.uk).

doi:10.1038/nature11467

Spontaneous giving and calculated greed

David G. Rand[1,2,3], Joshua D. Greene[2]* & Martin A. Nowak[1,4,5]*

Cooperation is central to human social behaviour[1–9]. However, choosing to cooperate requires individuals to incur a personal cost to benefit others. Here we explore the cognitive basis of cooperative decision-making in humans using a dual-process framework[10–18]. We ask whether people are predisposed towards selfishness, behaving cooperatively only through active self-control; or whether they are intuitively cooperative, with reflection and prospective reasoning favouring 'rational' self-interest. To investigate this issue, we perform ten studies using economic games. We find that across a range of experimental designs, subjects who reach their decisions more quickly are more cooperative. Furthermore, forcing subjects to decide quickly increases contributions, whereas instructing them to reflect and forcing them to decide slowly decreases contributions. Finally, an induction that primes subjects to trust their intuitions increases contributions compared with an induction that promotes greater reflection. To explain these results, we propose that cooperation is intuitive because cooperative heuristics are developed in daily life where cooperation is typically advantageous. We then validate predictions generated by this proposed mechanism. Our results provide convergent evidence that intuition supports cooperation in social dilemmas, and that reflection can undermine these cooperative impulses.

Many people are willing to make sacrifices for the common good[5–9]. Here we explore the cognitive mechanisms underlying this cooperative behaviour. We use a dual-process framework in which intuition and reflection interact to produce decisions[10–15,18]. Intuition is often associated with parallel processing, automaticity, effortlessness, lack of insight into the decision process and emotional influence. Reflection is often associated with serial processing, effortfulness and the rejection of emotional influence[10–15,18]. In addition, one of the psychological features most widely used to distinguish intuition from reflection is processing speed: intuitive responses are relatively fast, whereas reflective responses require additional time for deliberation[15]. Here we focus our attention on this particular dimension, which is closely related to the distinction between automatic and controlled processing[16,17].

Viewing cooperation from a dual-process perspective raises the following questions: are we intuitively self-interested, and is it only through reflection that we reject our selfish impulses and force ourselves to cooperate? Or are we intuitively cooperative, with reflection upon the logic of self-interest causing us to rein in our cooperative urges and instead act selfishly? Or, alternatively, is there no cognitive conflict between intuition and reflection? Here we address these questions using economic cooperation games.

We begin by examining subjects' decision times. The hypothesis that self-interest is intuitive, with prosociality requiring reflection to override one's selfish impulses, predicts that faster decisions will be less cooperative. Conversely, the hypothesis that intuition preferentially supports prosocial behaviour, whereas reflection leads to increased selfishness, predicts that faster decisions will be more cooperative.

As a first test of these competing hypotheses, we conducted a one-shot public goods game[5–8] (PGG) with groups of four participants.

We recruited 212 subjects from around the world using the online labour market Amazon Mechanical Turk (AMT)[19]. AMT provides a reliable subject pool that is more diverse than a typical sample of college undergraduates (see Supplementary Information, section 1). In accordance with standard AMT wages, each subject was given US$0.40 and was asked to choose how much to contribute to a common pool. Any money contributed was doubled and split evenly among the four group members (see Supplementary Information, section 3, for experimental details).

Figure 1a shows the fraction of the endowment contributed in the slower half of decisions compared to the faster half. Faster decisions result in substantially higher contributions compared with slower decisions (rank sum test, $P = 0.007$). Furthermore, as shown in Fig. 1b, we see a consistent decrease in contribution amount with

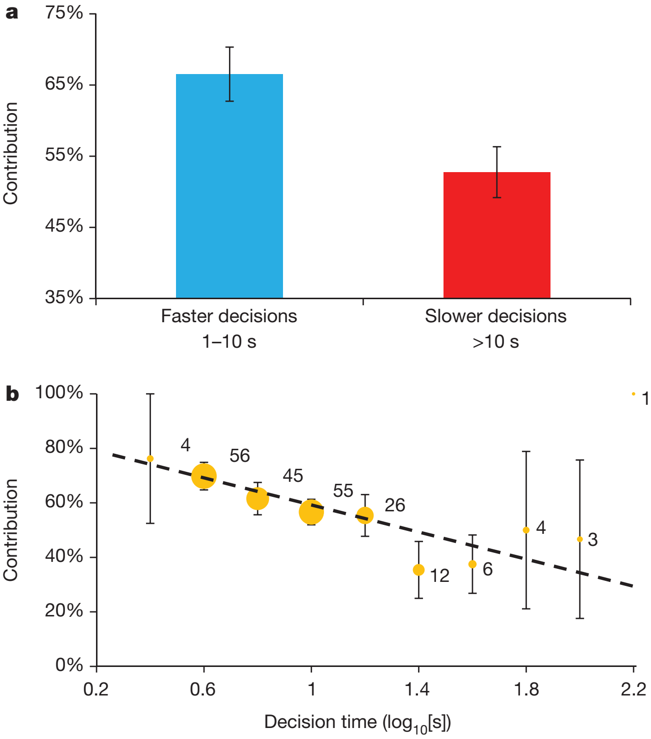


Figure 1 | Faster decisions are more cooperative. Subjects who reach their decisions more quickly contribute more in a one-shot PGG ($n = 212$). This suggests that the intuitive response is to be cooperative. **a**, Using a median split on decision time, we compare the contribution levels of the faster half versus slower half of decisions. The average contribution is substantially higher for the faster decisions. **b**, Plotting contribution as a function of $\log_{10}$-transformed decision time shows a negative relationship between decision time and contribution. Dot size is proportional to the number of observations, listed next to each dot. Error bars, mean ± s.e.m. (see Supplementary Information, sections 2 and 3, for statistical analysis and further details).

[1] Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts 02138, USA. [2] Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA. [3] Department of Psychology, Yale University, New Haven, Connecticut 06520, USA. [4] Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA. [5] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
*These authors contributed equally to this work.

increasing decision time (Tobit regression, coefficient = −15.84, $P = 0.019$; see Supplementary Information, sections 2 and 3, for statistical details). These findings suggest that intuitive responses are more cooperative.

Next we examined data from all of our previously published social dilemma experiments for which decision time data were recorded[7,20–22]. In these studies, conducted in the physical laboratory with college students, the experimental software automatically recorded decision times, but these data had not been previously analysed. To examine the psychology that subjects bring with them into the laboratory, we focused on play in the first round of each experimental session. In a one-shot prisoner's dilemma ($n = 48$)[20], a repeated prisoner's dilemma with execution errors ($n = 278$)[21], a repeated prisoner's dilemma with and without costly punishment ($n = 104$)[22], and a repeated PGG with and without reward and/or punishment ($n = 192$)[7], we find the same negative relationship between decision time and cooperation (see Supplementary Information, section 4, for details). These results show the robustness of our decision-time findings: across a range of experimental designs, and with students in the physical laboratory as well as with an international online sample, faster decisions are associated with more prosociality.

We now demonstrate the causal link between intuition and cooperation suggested by these correlational studies. To do so, we recruited another 680 subjects on AMT and experimentally manipulated their decision times in the same one-shot PGG used above. In the 'time pressure' condition, subjects were forced to reach their decision quickly (within 10 s). Subjects in this condition have less time to reflect than in a standard PGG, and therefore their decisions are expected to be more intuitive. In the 'time delay' condition, subjects were instructed to carefully consider their decision and forced to wait for at least 10 s before choosing a contribution amount. Thus, in this condition, decisions are expected to be driven more by reflection (see Supplementary Information, section 5, for experimental details).

The results (Fig. 2a) are consistent with the correlational observations in Fig 1. Subjects in the time-pressure condition contribute significantly more money on average than subjects in the time-delay condition (rank sum, $P < 0.001$). Moreover, we find that both manipulation conditions differ from the average behaviour in the baseline experiment in Fig. 1, and in the expected directions: subjects under time-pressure contribute more than unconstrained subjects (rank sum, $P = 0.058$), whereas subjects who are instructed to reflect and delay their decision contribute less than unconstrained subjects (rank sum, $P = 0.028$), although the former difference is only marginally significant. See Supplementary Information, section 5, for regression analyses.

Additionally, we recruited 211 Boston-area college students and replicated our time-constraint experiment in the physical laboratory with tenfold higher stakes (Fig. 2b). We find again that subjects in the time-pressure condition contribute significantly more money than subjects in the time-delay condition (rank sum, $P = 0.032$). We also assessed subjects' expectations about the behaviour of others in their group, and find no significant difference across conditions (rank sum, $P = 0.360$). Thus, subjects forced to respond more intuitively seem to have more prosocial preferences, rather than simply contributing more because they are more optimistic about the behaviour of others (see Supplementary Information, section 6, for experimental details and analysis).

We next used a conceptual priming manipulation that explicitly invokes intuition and reflection[23]. We recruited 343 subjects on AMT to participate in a one-shot PGG experiment. The first condition promotes intuition relative to reflection: before reading the PGG instructions, subjects were assigned to write a paragraph about a situation in which either their intuition had led them in the right direction, or careful reasoning had led them in the wrong direction. Conversely, the second condition promotes reflection: subjects were asked to write about either a situation in which intuition led them in the wrong direction, or careful reasoning had led them in the right direction. Consistent with the seven experiments described above, we find that contributions are significantly higher when subjects are primed to promote intuition relative to reflection (Fig. 2c; rank sum, $P = 0.011$; see Supplementary Information, section 8, for experimental details and analysis).

These results therefore raise the question of why people are intuitively predisposed towards cooperation. We propose the following mechanism: people develop their intuitions in the context of daily life, where cooperation is typically advantageous because many important interactions are repeated[1,2,21,22], reputation is often at

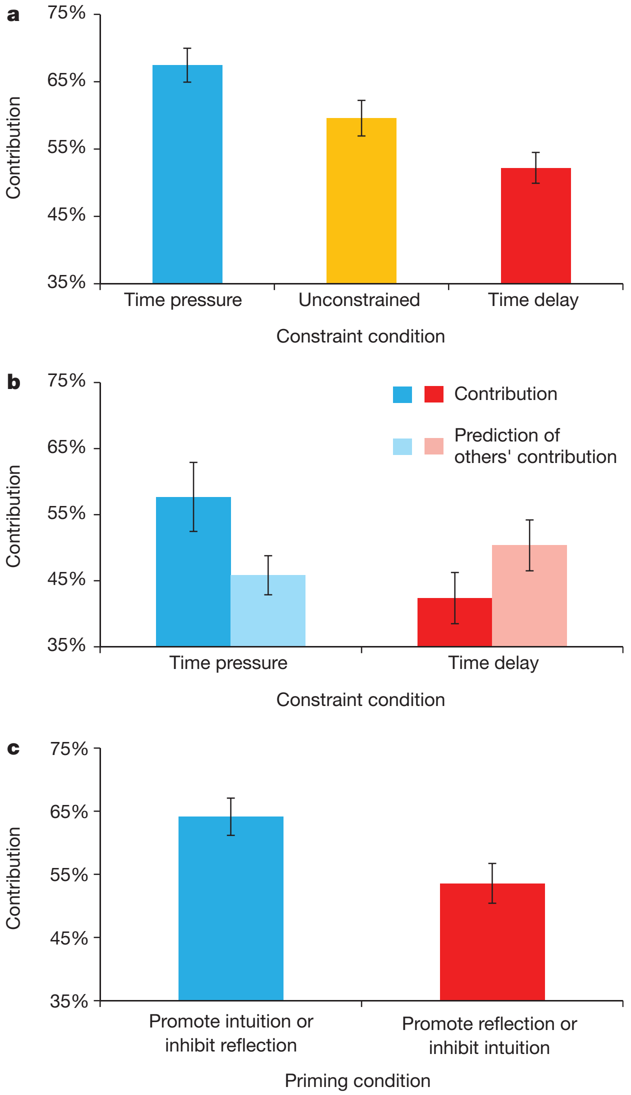


Figure 2 | Inducing intuitive thinking promotes cooperation. **a**, Forcing subjects to decide quickly (10 s or less) results in higher contributions, whereas forcing subjects to decide slowly (more than 10 s) decreases contributions ($n = 680$). This demonstrates the causal link between decision time and cooperation suggested by the correlation shown in Fig. 1. **b**, We replicate the finding that forcing subjects to decide quickly promotes cooperation in a second study run in the physical laboratory with tenfold larger stakes ($n = 211$). We also find that the time constraint has no significant effect on subjects' predictions concerning the average contributions of other group members. Thus, the manipulation acts through preferences rather than beliefs. **c**, Priming intuition (or inhibiting reflection) increases cooperation relative to priming reflection (or inhibiting intuition) ($n = 343$). This finding provides further evidence for the specific role of intuition versus reflection in motivating cooperation, as suggested by the decision time studies. Error bars, mean ± s.e.m. (see Supplementary Information, sections 5–7, for statistical analysis and further details).

stake[3,5,6,20] and sanctions for good or bad behaviour might exist[4,6–8]. Thus, our subjects develop cooperative intuitions for social interactions and bring these cooperative intuitions with them into the laboratory. As a result, their automatic first response is to be cooperative. It then requires reflection to overcome this cooperative impulse and instead adapt to the unusual situation created in these experiments, in which cooperation is not advantageous.

This hypothesis makes clear predictions about individual difference moderators of the effect of intuition on cooperation, two of which we now test. First, if the effects described above result from intuitions formed through ordinary experience, then greater familiarity with laboratory cooperation experiments should attenuate these effects. We test this prediction on AMT with a replication of our conceptual priming experiment. As predicted, we find a significant interaction between prime and experience: it is only among subjects naive to the experimental task that promoting intuition increases cooperation (Fig. 3a; see Supplementary Information, section 9, for experimental details and statistical analysis).

This mechanism also predicts that subjects will only find cooperation intuitive if they developed their intuitions in daily-life settings in which cooperation was advantageous. Even in the presence of repetition, reputation and sanctions, cooperation will only be favoured if enough other people are similarly cooperative[2,3]. We tested this prediction on AMT with a replication of our baseline correlational study. As predicted, it is only among subjects that report having mainly cooperative daily-life interaction partners that faster decisions are associated with higher contributions (Fig. 3b; see Supplementary Information, section 10, for experimental details and statistical analysis).

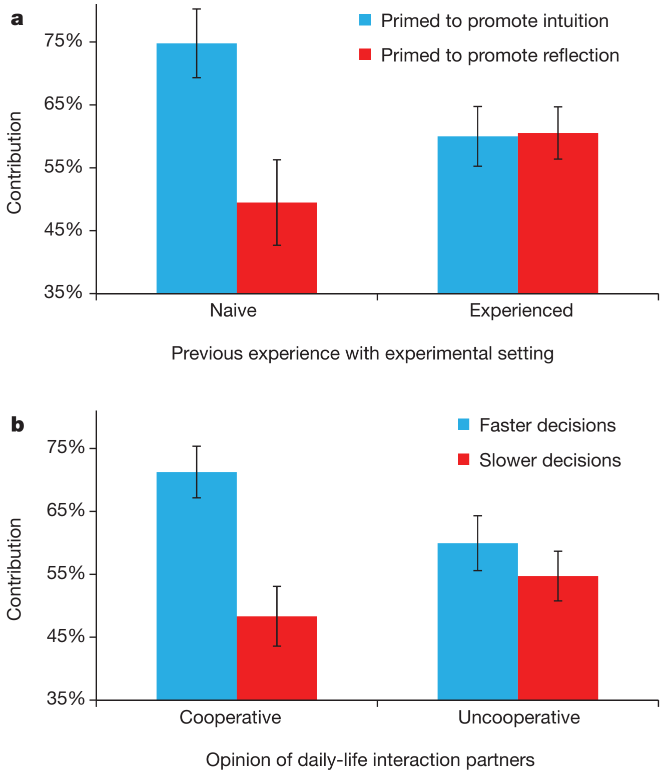


Figure 3 | Evidence that cooperative intuitions from daily lift spill over into the laboratory. Two experiments validate predictions of our hypothesis that subjects develop their cooperative intuitions in the context of daily life, in which cooperation is advantageous. **a**, Priming that promotes reliance on intuition increases cooperation relative to priming promoting reflection, but only among naive subjects that report no previous experience with the experimental setting where cooperation is disadvantageous ($n = 256$). **b**, Faster decisions are associated with higher contribution levels, but only among subjects who report having cooperative daily-life interaction partners ($n = 341$). As in Fig. 1a, a median split is carried out on decision times, separating decisions into the faster versus slower half. Error bars, mean ± s.e.m. (see Supplementary Information, sections 9 and 10, for statistical analysis and further details).

Thus, there are some people for whom the intuitive response is more cooperative and the reflective response is less cooperative; and there are other people for whom both the intuitive and reflective responses lead to relatively little cooperation. But we find no cases in which the intuitive response is reliably less cooperative than the reflective response. As a result, on average, intuition promotes cooperation relative to reflection in our experiments.

By showing that people do not have a single consistent set of social preferences, our results highlight the need for more cognitively complex economic and evolutionary models of cooperation, along the lines of recent models for non-social decision-making[17,24–26]. Furthermore, our results suggest a special role for intuition in promoting cooperation[27]. For further discussion, and a discussion of previous work exploring behaviour in economic games from a dual-process perspective, see Supplementary Information, sections 12 and 13.

On the basis of our results, it may be tempting to conclude that cooperation is 'innate' and genetically hardwired, rather than the product of cultural transmission. This is not necessarily the case: intuitive responses could also be shaped by cultural evolution[28] and social learning over the course of development. However, our results are consistent with work demonstrating spontaneous helping behaviour in young children[29]. Exploring the role of intuition and reflection in cooperation among children, as well as cross-culturally, can shed further light on this issue.

Here we have explored the cognitive underpinnings of cooperation in humans. Our results help to explain the origins of cooperative behaviour, and have implications for the design of institutions that aim to promote cooperation. Encouraging decision-makers to be maximally rational may have the unintended side-effect of making them more selfish. Furthermore, rational arguments about the importance of cooperating may paradoxically have a similar effect, whereas interventions targeting prosocial intuitions may be more successful[30]. Exploring the implications of our findings, both for scientific understanding and public policy, is an important direction for future study: although the cold logic of self-interest is seductive, our first impulse is to cooperate.

METHODS SUMMARY

Across studies 1, 6, 8, 9 and 10, a total of 1,955 subjects were recruited using AMT[19] to participate in one of a series of variations on the one-shot PGG, played through an online survey website. Subjects received $0.50 for participating, and could earn up to $1 more based on the PGG. In the PGG, subject were given $0.40 and chose how much to contribute to a 'common project'. All contributions were doubled and split equally among four group members. Once all subjects in the experiment had made their decisions, groups of four were randomly matched and the resulting payoffs were calculated. Each subject was then paid accordingly through the AMT payment system, and was informed about the average contribution of the other members of his or her group. No deception was used.

In study 7, a total of 211 subjects were recruited from the Boston, Massachusetts, metropolitan area through the Harvard University Computer Laboratory for Experiment Research subject pool to participate in an experiment at the Harvard Decision Science Laboratory. Participation was restricted to students under 35 years of age. Subjects received a $5 show-up fee for arriving on time and had the opportunity to earn up to an additional $12 in the experiment. Subjects played a single one-shot PGG through the same website interface used in the AMT studies, but with tenfold larger stakes (maximum earnings of $10). Subjects were then asked to predict the average contribution of their other group members and had the chance to win up to an additional $2 based on their accuracy.

These experiments were approved by the Harvard University Committee on the Use of Human Subjects in Research.

For further details of the experimental methods, see Supplementary Information.

Received 13 December 2011; accepted 2 August 2012.

1. Trivers, R. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46,** 35–57 (1971).

2. Fudenberg, D. & Maskin, E. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54,** 533–554 (1986).
3. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437,** 1291–1298 (2005).
4. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100,** 3531–3535 (2003).
5. Milinski, M., Semmann, D. & Krambeck, H. J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415,** 424–426 (2002).
6. Rockenbach, B. & Milinski, M. The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444,** 718–723 (2006).
7. Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M. A. Positive interactions promote public cooperation. *Science* **325,** 1272–1275 (2009).
8. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415,** 137–140 (2002).
9. Rand, D. G., Arbesman, S. & Christakis, N. A. Dynamic social networks promote cooperation in experiments with humans. *Proc. Natl Acad. Sci. USA* **108,** 19193–19198 (2011).
10. Sloman, S. A. The empirical case for two systems of reasoning. *Psychol. Bull.* **119,** 3–22 (1996).
11. Stanovich, K. E. & West, R. F. Individual differences in rational thought. *J. Exp. Psychol.* **127,** 161–188 (1998).
12. Chaiken, S. & Trope, Y. *Dual-Process Theories in Social Psychology* (Guilford, 1999).
13. Kahneman, D. A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* **58,** 697–720 (2003).
14. Plessner, H., Betsch, C. & Betsch, T. *Intuition in Judgment and Decision Making* (Lawrence Erlbaum, 2008).
15. Kahneman, D. *Thinking, Fast and Slow* (Straus and Giroux, 2011).
16. Shiffrin, R. M. & Schneider, W. Controlled and automatic information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychol. Rev.* **84,** 127–190 (1977).
17. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24,** 167–202 (2001).
18. Frederick, S. Cognitive reflection and decision making. *J. Econ. Perspect.* **19,** 25–42 (2005).
19. Horton, J. J., Rand, D. G. & Zeckhauser, R. J. The online laboratory: conducting experiments in a real labor market. *Exp. Econ.* **14,** 399–425 (2011).
20. Pfeiffer, T., Tran, L., Krumme, C. & Rand, D. G. The value of reputation. *J. R. Soc. Interface* http://dx.doi.org/10.1098/rsif.2012.0332 (20 June 2012).
21. Fudenberg, D., Rand, D. G. & Dreber, A. Slow to anger and fast to forgive: cooperation in an uncertain world. *Am. Econ. Rev.* **102,** 720–749 (2012).
22. Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. A. Winners don't punish. *Nature* **452,** 348–351 (2008).
23. Shenhav, A., Rand, D. G. & Greene, J. D. Divine intuition: cognitive style influences belief in God. *J. Exp. Psychol. Gen.* **141,** 423–428 (2012).
24. Benhabib, J. & Bisin, A. Modeling internal commitment mechanisms and self-control: a neuroeconomics approach to consumption–saving decisions. *Games Econ. Behav.* **52,** 460–492 (2005).
25. Fudenberg, D. & Levine, D. K. A. Dual-self model of impulse control. *Am. Econ. Rev.* **96,** 1449–1476 (2006).
26. McClure, S. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. Separate neural systems value immediate and delayed monetary rewards. *Science* **306,** 503–507 (2004).
27. Bowles, S. & Gintis, H. in *The Economy as a Evolving Complex System 3* (eds Blume, L. and Durlauf, S. N.) 339–364 (2002).
28. Richerson, P. J. & Boyd, R. *Not by Genes Alone: How Culture Transformed Human Evolution*. (Univ. Chicago Press, 2005).
29. Warneken, F. & Tomasello, M. Altruistic helping in human infants and young chimpanzees. *Science* **311,** 1301–1303 (2006).
30. Bowles, S. Policies designed for self-interested citizens may undermine "the moral sentiments": evidence from economic experiments. *Science* **320,** 1605–1609 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank H. Ahlblad, O. Amir, F. Fu, O. Hauser, J. Horton and R. Kane for assistance with carrying out the experiments, and P. Blake, S. Bowles, N. Christakis, F. Cushman, A. Dreber, T. Ellingsen, F. Fu, D. Fudenberg, O. Hauser, J. Jordan, M. Johannesson, M. Manapat, J. Paxton, A. Peysakhovich, A. Shenhav, J. Sirlin-Rand, M. van Veelen and O. Wurzbacher for discussion and comments. This work was supported in part by a National Science Foundation grant (SES-0821978 to J.D.G.). D.G.R. and M.A.N. are supported by grants from the John Templeton Foundation.

Author Contributions D.G.R., J.D.G. and M.A.N. designed the experiments, D.G.R. carried out the experiments and statistical analyses, and D.G.R., J.D.G. and M.A.N. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.G.R. (drand@fas.harvard.edu).

doi:10.1038/nature11330

Sex-specific volatile compounds influence microarthropod-mediated fertilization of moss

Todd N. Rosenstiel[1], Erin E. Shortlidge[1], Andrea N. Melnychenko[1], James F. Pankow[2,3] & Sarah M. Eppley[1]

Sexual reproduction in non-vascular plants requires unicellular free-motile sperm to travel from male to female reproductive structures across the terrestrial landscape[1]. Recent data suggest that microarthropods can disperse sperm in mosses[2]. However, little is known about the chemical communication, if any, that is involved in this interaction or the relative importance of microarthropod dispersal compared to abiotic dispersal agents in mosses. Here we show that tissues of the cosmopolitan moss *Ceratodon purpureus* emit complex volatile scents, similar in chemical diversity to those described in pollination mutualisms between flowering plants and insects, that the chemical composition of *C. purpureus* volatiles are sex-specific, and that moss-dwelling microarthropods are differentially attracted to these sex-specific moss volatile cues. Furthermore, using experimental microcosms, we show that microarthropods significantly increase moss fertilization rates, even in the presence of water spray, highlighting the important role of microarthropod dispersal in contributing to moss mating success. Taken together, our results indicate the presence of a scent-based 'plant–pollinator-like' relationship that has evolved between two of Earth's most ancient terrestrial lineages, mosses and microarthropods.

The origin of bryophytes (mosses, liverworts and hornworts) during the upper Ordovician period represents a notable event in the evolution of life[3,4], leading to the diversification of terrestrial organisms. From a mating systems perspective, the evolution of bryophytes resulted in sexual reproduction partially escaping the aquatic environment. In mosses, sexual reproduction requires free-motile sperm to 'swim' with the aid of water across the terrestrial landscape to fertile females[1,5], a reminder of its aquatic origins. This model of 'swimming sperm' has led to the general view that sperm dispersal among bryophytes is quite limited, with most fertilization occurring within about 10 cm (refs 6, 7). However, recent research using the moss *Bryum argenteum* shows that moss sperm can be dispersed by microarthropods[2], specifically springtails and oribatid mites, which are common inhabitants of moss patches worldwide[8]. This new research builds on earlier, often overlooked work indicating that arthropods may act as ecologically relevant sperm transport vectors[5]. Furthermore, recent data show that moss sperm can be more long-lived and stress tolerant than believed previously[9,10], potentially enabling sperm to survive during long-distance microarthropod dispersal. Although microarthropods are known to use volatile cues in foraging and for communication[11-13], little is known about whether microarthropods may also use chemical cues to facilitate sexual reproduction in mosses.

Here we assess the potential role of moss volatile cues and microarthropods (springtails) in mediating sperm dispersal in *C. purpureus*, a model cosmopolitan moss species with separate sexes. First, to fully capture the suite of possible volatile organic compounds (VOCs) emitted from intact (non-wounded), sexually expressing (gametoecia-producing) male and female plants, we characterized headspace VOCs using two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC–TOFMS). We found that for all sampled populations, female plants released a significantly greater number of VOCs than male plants (104.00 ± 9.27 and 29.86 ± 8.21, respectively; $P < 0.0001$; Fig. 1). In addition, analyses of VOC composition revealed significant sex-specific differences (analysis of similarities (ANOSIM): $R = 0.79$, $P = 0.001$, stress value = 3.8; Fig. 2). A surprising diversity of volatile compounds was identified in headspace analysis using our GC × GC–TOFMS approach, and many of these compounds have been identified previously in floral scents of flowering plants[14].

Second, to determine whether springtails were differentially attracted to the observed sex-specific VOC composition, we conducted a series of preference assays using intact (non-wounded) samples of male and female *C. purpureus*. In the first set of preference assays, springtails were given choices between male and female moss samples in Petri dishes, and were found to be significantly more likely to choose intact reproductive female plants over intact reproductive male plants

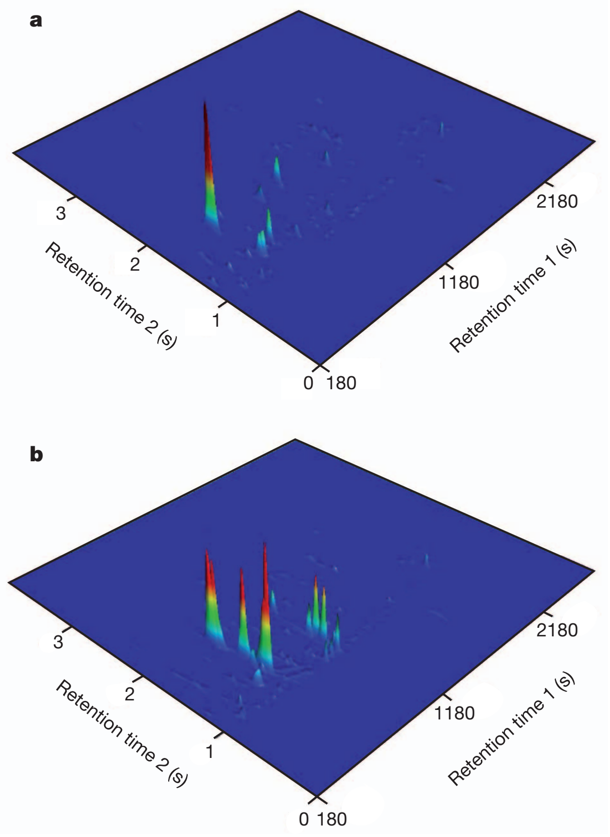


Figure 1 | Sex-specific volatile profiles. a, b, Representative two-dimensional GC × GC–TOFMS chromatograms of volatile compounds from intact shoots of a reproductive male (**a**) and a reproductive female (**b**) of the cosmopolitan moss *C. purpureus*. Colours indicate relative measures of compound abundance; red indicates compounds that are greater than 50% of the largest individual peak area.

[1]Department of Biology and Center for Life in Extreme Environments, Portland State University, 1719 SW 10th Avenue, Portland, Oregon 97201, USA. [2]Department of Chemistry, 1719 SW 10th Avenue, Portland State University, Portland, Oregon 97201, USA. [3]Department of Civil and Environmental Engineering, Portland State University, 1930 SW 4th Avenue, Portland, Oregon 97201, USA.

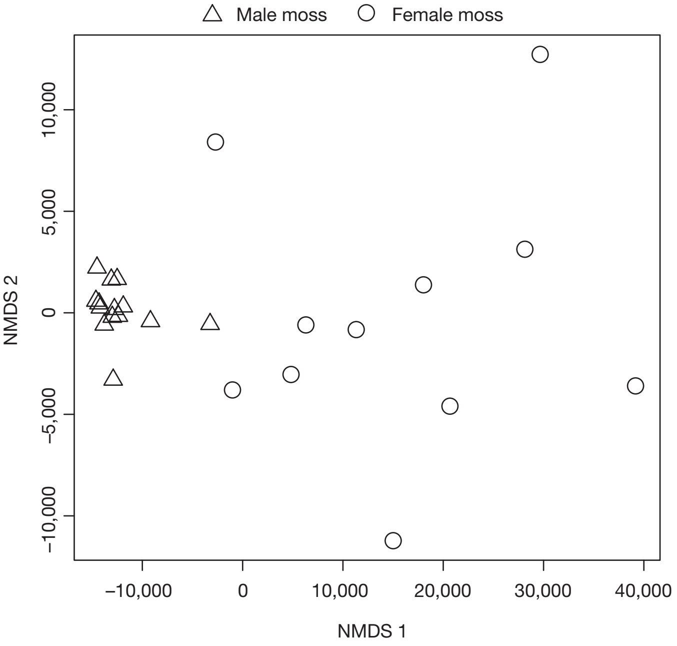


Figure 2 | Differences in volatile composition. Non-metric multidimensional scaling (NMDS) of volatile scent profiles of reproductive male and reproductive female plants of *C. purpureus* shows that there are significant sex-specific differences in VOC composition ($P = 0.001$). Symbols represent scent profiles of individual males and females ($n = 22$, GC × GC–TOFMS analyses).

($G = 37.6$; $P < 0.0001$; Fig. 3). This result is similar to the that of another study[2], in which springtails and mites marginally preferred female to male reproductive *B. argenteum* plants. To confirm that springtail preference for female plants was due to female-specific volatile cues, we used an olfactometer for additional preference assays. Microarthropods were able to assess scents produced by the samples but were not given access to visual or physical cues. In the olfactometer assays, springtails chose intact reproductive female plants significantly more frequently than intact reproductive male plants ($G = 58.1$, $P < 0.0001$). These results reveal the surprising role of volatile cues in influencing microarthropods' choice of intact female moss plants.

Last, we used a series of microcosm experiments in which we manipulated springtail abundance and water spray to assess the importance of biotic versus abiotic factors in promoting sperm dispersal and fertilization in mosses. For this experiment, we used *C. purpureus* and *B. argenteum* (this is the moss species for which springtail-mediated sperm dispersal has been demonstrated previously)[2]. Our results show that for both moss species, the addition of either springtails or water spray significantly increased the number of sporophytes formed per microcosm ($P = 0.05$ and $P = 0.02$, respectively) and the fraction of microcosms that developed sporophytes ($P = 0.03$ and $P = 0.03$, respectively; Fig. 4). Moreover, the combination of treatments had a pronounced synergistic effect, more than doubling the effect of either treatment alone ($P = 0.03$ for the number of sporophytes per microcosm, and $P = 0.03$ for the fraction of microcosms with sporophytes; Fig. 4c). These results highlight the substantial role of microarthropods in facilitating fertilization in mosses, presumably through enhanced sperm transport.

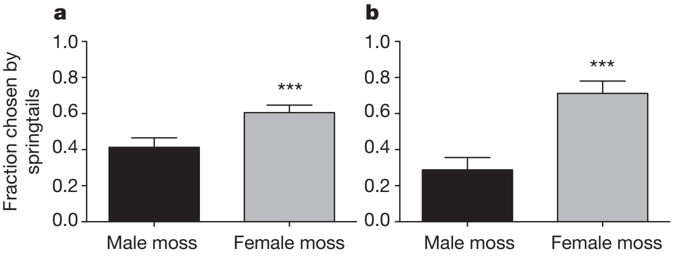


Figure 3 | Springtails prefer female moss. **a**, **b**, The fraction of *C. purpureus* samples chosen by springtails (error bars, mean ± s.e.m.) in preference assays of male versus female samples in Petri dishes (**a**, $n = 24$ assays, 491 springtails counted); and male versus female samples in an olfactometer (**b**, $n = 10$ assays, 276 springtails counted). ***$P < 0.0001$.

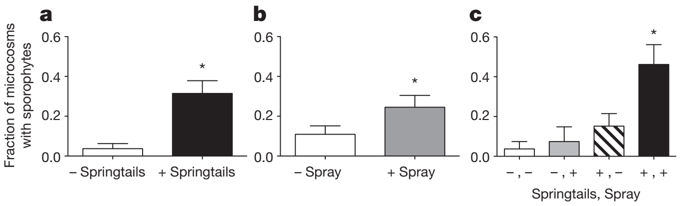


Figure 4 | Springtails enhance fertilization in moss microcosms. **a**–**c**, Fertilization success in *C. purpureus* and *B. argenteum* microcosms, measured as the fraction of microcosms that developed sporophytes (error bars, mean ± s.e.m.). The effects of springtail treatment (**a**), water spray treatment (**b**) and the interaction between these treatments (**c**) on fertilization success. Plus and minus symbols represent the presence and absence of springtails and water spray. $n = 108$ microcosms. *$P < 0.05$.

Plant–insect interactions were key to the diversification of flowering plants[15], with floral scent representing a primary mode of communication between plants and their pollinators[16,17]. Our data suggest that mosses, despite their lack of flowering structures, may similarly utilize volatile scents as cues to manipulate microarthropod behaviour, resulting in increased moss fertilization. Therefore, we propose that there may be a notable plant–pollinator-like relationship that has evolved between microarthropods and mosses involving volatile scent cues.

Sex-specific floral scents have been found in over 20 species of flowering plants with separate sexes, and several ideas have been proposed to explain this pattern[18]. One idea is that the most mate-limited sex is likely to evolve the greatest floral scent[18]. If this theory extends to bryophytes, then our results suggest that female mosses are more mate-limited than males, which is likely given the highly female-biased population sex ratios of these species[19,20], as is typical in mosses[21]. Another idea is that differential pollinator rewards between the sexes may lead to selection for differential cues, including sex-specific VOCs[22]. If, during the normal course of their movements, microarthropods inadvertently pick up released moss sperm from water film[5], or if moss sperm are a food reward for microarthropods (similar to pollen in some plant-pollinator systems), then the reward and cues for male and female moss plants are likely to be different. For example, it has been suggested that females may produce high concentrations of sucrose or fatty acids as a reward[2]. We have not yet distinguished between the composition and amounts of VOCs produced by the reproductive structures and the entire plant, or between the moss tissue and any associated phyllospheric microbes. Sex-specific mutualistic interactions do occur between hosts and microbes[23] and can induce sex-specific VOC differences in the host[24], and it is possible that such interactions may exist in bryophytes. Further studies are needed to establish the fundamental factors in this moss–microarthropod signalling system, including determining which specific VOCs, or suites of VOCs, are most important for signalling as well as pinpointing the cells responsible for the production of key volatile cues. As mosses and microarthropods are two of Earth's most ancient co-occurring terrestrial lineages, it is important to consider the potential role that a plant–pollinator-like relationship may have had in shaping the evolutionary ecology of moss mating systems.

METHODS SUMMARY

To examine volatile profiles in these mosses, gas chromatography was carried out using a Pegasus 4D GC × GC–TOFMS system (LECO). For each sample, 30–40 mg of intact (non-wounded) moss shoots was allowed to equilibrate in a glass vial for 120 min. Headspace sampling was carried out for 60 min with a solid phase microextraction (SPME) fibre, then thermal desorption of the SPME fibre and analysis by GC × GC-TOFMS was carried out as described previously[25]. Data are based on *C. purpureus* plants collected in Oregon and maintained in greenhouse culture.

To determine springtail preference for male versus female *C. purpureus* samples of intact shoots, we conducted two sets of preference assays. First, for preference assays of whole moss patches, protocols were modified from well-established springtail food preference assays in Petri dishes[26,27], and we used *C. purpureus* plants collected in Oregon and maintained in greenhouse culture. Second, for volatile preference assays, we used a custom-constructed static-air olfactometer designed for springtails[28], and *C. purpureus* plants were collected directly from the field in Oregon. We used two springtail species, *Folsomia candida* and *Sinella curviseta*, for both sets of assays.

To determine the effect of springtails and water spray on moss fertilization, we maintained microcosms of *C. purpureus* ($n = 72$ microcosms) and *B. argenteum* ($n = 36$ microcosms) for approximately 15 weeks in a factorial design with treatments of added springtails and water spray, counting the number of sporophytes after initial sporophyte formation. *C. purpureus* and *B. argenteum* plants were collected in Oregon, Arizona and Kentucky, and maintained in greenhouse culture.

Full Methods and any associated references are available in the online version of the paper.

Received 16 May; accepted 15 June 2012.
Published online 18 July 2012.

1. Paolillo, D. J. J. The swimming sperms of land plants. *Bioscience* **31,** 367–373 (1981).
2. Cronberg, N., Natcheva, R. & Hedlund, K. Microarthropods mediate sperm transfer in mosses. *Science* **313,** 1255 (2006).
3. Nickrent, D. L., Parkinson, C. L., Palmer, J. D. & Duff, R. J. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **17,** 1885–1895 (2000).
4. Kenrick, P. & Crane, P. R. The origin and early evolution of land plants. *Nature* **389,** 33–39 (1997).
5. Muggoch, H. & Walton, J. On the dehiscence of the antheridium and the part played by surface tension in the dispersal of spermatocytes in Bryophyta. *Proc. R.. Soc. Lond. B* **130,** 448–461 (1942).
6. Longton, R. E. Reproductive biology and evolutionary potential in bryophytes. *J. Hattori Bot. Lab.* **41,** 205–223 (1976).
7. Shaw, A. J. in *Bryophyte Biology* (eds Shaw, A. J. & Goffinet, B.) 369–402 (Cambridge Univ. Press, 2000).
8. Andrew, N. R., Rodgerson, L. & Dunlop, M. Variation in invertebrate-bryophyte community structure at different spatial scales along altitudinal gradients. *J. Biogeogr.* **30,** 731–746 (2003).
9. Rosenstiel, T. N. & Eppley, S. M. Long-lived sperm in the geothermal bryophyte *Pohlia nutans*. *Biol. Lett.* **5,** 857–860 (2009).
10. Shortlidge, E. E., Rosenstiel, T. N. & Eppley, S. M. Tolerance to environmental desiccation in moss sperm. *New Phytol.* **194,** 741–750 (2012).
11. Verhoef, H. A., Nagelkerke, C. J. & Joosse, E. N. G. Aggregation pheromones in Collembola. *J. Insect Physiol.* **23,** 1009–1013 (1977).
12. Raspotnig, G., Krisper, G., Schuster, R., Fauler, G. & Leis, H. J. Volatile exudates from the oribatid mite, *Platynothrus peltifer*. *J. Chem. Ecol.* **31,** 419–430 (2005).
13. Bengtsson, G., Erlandsson, A. & Rundgren, S. Fungal odour attracts soil Collembola. *Soil Biol. Biochem.* **20,** 25–30 (1988).
14. Knudsen, J. T., Tollsten, L. & Bergstrom, L. G. Floral scents — a checklist of volatile compounds isolated by headspace techniques. *Phytochemistry* **33,** 253–280 (1993).
15. Crepet, W. L. Advanced (constant) insect pollination mechanisms: pattern of evolution and implications vis-a-vis angiosperm diversity. *Ann. Mo. Bot. Gard.* **71,** 607–630 (1984).
16. Schiestl, F. P. The evolution of floral scent and insect chemical communication. *Ecol. Lett.* **13,** 643–656 (2010).
17. Raguso, R. A. Wake up and smell the roses: the ecology and evolution of floral scent. *Annu. Rev. Ecol. Evol. Syst.* **39,** 549–569 (2008).
18. Ashman, T. L. Sniffing out patterns of sexual dimorphism in floral scent. *Funct. Ecol.* **23,** 852–862 (2009).
19. Shaw, A. J. & Gaughan, J. F. Control of sex-ratios in haploid populations of the moss, *Ceratodon purpureus*. *Am. J. Bot.* **80,** 584–591 (1993).
20. Stark, L. R., McLetchie, D. N. & Eppley, S. M. Sex ratios and the shy male hypothesis in *Bryum argenteum* (Bryaceae). *Bryologist* **113,** 788–797 (2010).
21. Bisang, I. & Hedenäs, L. Sex ratio patterns in dioicous bryophytes re-visited. *J. Bryol.* **27,** 207–219 (2005).
22. Hemborg, A. M. & Bond, W. J. Different rewards in female and male flowers can explain the evolution of sexual dimorphism in plants. *Biol. J. Linn. Soc.* **85,** 97–109 (2005).
23. Varga, S. & Kytoviita, M. M. Sex-specific responses to mycorrhiza in a dioecious species. *Am. J. Bot.* **95,** 1225–1232 (2008).
24. Voigt, C. C., Caspers, B. & Speck, S. Bats, bacteria, and bat smell: sex-specific diversity of microbes in a sexually selected scent organ. *J. Mamm.* **86,** 745–749 (2005).
25. Pankow, J. F. *et al.* Volatilizable biogenic organic compounds (VBOCs) with two dimensional gas chromatography-time of flight mass spectrometry (GC×GC–TOFMS): sampling methods, VBOC complexity, and chromatographic retention data. *Atmos. Meas. Tech. Discuss.* **4,** 3647–3684 (2011).
26. Thimm, T. & Larink, O. Grazing preferences of some collembola for endomycorrhizal fungi. *Biol. Fertil. Soils* **19,** 266–268 (1995).
27. Sadaka-Laulan, N., Ponge, J. F., Roquebert, M. F., Bury, E. & Boumezzough, A. Feeding preferences of the Collembolan *Onychiurus sinensis* for fungi colonizing holm oak litter (*Quercus rotundifolia* Lam.). *Eur. J. Soil Biol.* **34,** 179–188 (1998).
28. Staaden, S., Milcu, A., Rohlfs, M. & Scheu, S. Olfactory cues associated with fungal grazing intensity and secondary metabolite pathway modulate Collembola foraging behaviour. *Soil Biol. Biochem.* **43,** 1411–1416 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to W. Luo and L. Isabelle for analytical assistance, to C. Rupert for assistance with the preference assays, and to L. Stark and N. McLetchie for plant material. Funding was provided by the 3M Corporation and the National Science Foundation (DEB-0743461 to S.M.E. and IOS-0719570 to T.N.R).

Author Contributions S.M.E. and T.N.R. coordinated and planned the project. E.E.S., S.M.E., and T.N.R. initiated and carried out the experiments. A.N.M., J.F.P.. and T.N.R. were responsible for the gas chromatography data and analyses. All authors contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.M.E. (eppley@pdx.edu).

METHODS

Study system. *Ceratodon purpureus* (Hedw.) Brid. and *Bryum argenteum* Hedw. are nearly cosmopolitan species, with dioicous breeding systems[29]. For this study, in 2009, we collected plants from three *C. purpureus* populations in the Portland, Oregon metro area from northeast Portland, the Portland State University campus and a farm in North Plains, Oregon, with populations a minimum of 5.8 km apart. The *B. argenteum* plants were collected from 2008 to 2009, from four populations from southern Arizona, the University of Kentucky campus, southwest Portland, Oregon and downtown Portland, Oregon, with populations a minimum of 1.9 km apart. Plants were collected from field populations and grown in the Portland State University greenhouse in pots (6.4 × 6.4 cm) for at least 3 months for all experiments except the olfactometer experiment (for which plants were collected directly from the field). Plants from the *C. purpureus* northeast Portland and Portland State University populations, and the *B. argenteum* Arizona population were grown from spores (single-spore isolations) to ensure that there were separate individuals from these smaller populations, whereas the other plants were grown from single-shoot cuttings, and plants as far apart as possible were collected from within populations. Two commercially available species of springtails, *Folsomia candida* and *Sinella curviseta*, were reared in airtight containers with natural charcoal, de-ionized water and yeast, and were kept in the same growth chambers in which the microcosm experiments were performed (see below). *F. candida* is a model springtail species, growing in soils worldwide[30], and it occurs in high densities in soil and moss communities in the Pacific Northwest region of the Unites States and Canada[31]. *S. curviseta* is an emerging model system as it occurs in sites where *F. candida* is rare[32]. Both species were used in all springtail experiments.

Scent collection and GC × GC–TOFMS analyses. To examine the volatile scent profiles of intact moss tissue, we used static headspace, a method that is sensitive to and can therefore identify small quantities of compounds but cannot be easily used for quantification of amounts of compounds in volatile signatures. For each tissue sample, 30–40 mg of intact (non-wounded) shoots was carefully removed from pots. Each sample was placed into a 2-ml screw-top glass vial and allowed to equilibrate for 120 min. A solid phase micro-extraction (SPME) fibre (polydimethylsiloxane–divinylbenzene, 65 μm coating; Sigma-Aldrich) was then exposed to the headspace for an additional 60 min; results did not change appreciably with additional exposure time (data not shown). Each analysis began by inserting the SPME fibre into the injector (with 'SPME liner') of a two-dimensional gas chromatograph (Pegasus 4D GC × GC–TOFMS system; LECO); the column and analytical conditions used were as described previously for biogenic volatiles[25]. Trace contaminants from ambient air blanks were identified and removed from each of the comprehensive volatile profiles before further data analysis. Comprehensive GC × GC–TOFMS analysis was chosen to minimize any a priori assumptions about the chemical nature of the volatile compounds emitted by the moss system.

To determine whether male and female plants of *C. purpureus* differed in scent composition, we compared overall variation in chemical composition between the sexes, using males and females from each of the three Portland, Oregon populations ($n = 22$ plants in total). For all analyses, we used plants that were producing gametoecia (perichaetia and perigonium, female and male sex organs, respectively, with clusters of modified leaves), and we enriched these structures in the samples. However, initial screens of plant material without gametoecia suggest that male and female plants show a similar difference in volatile composition (data not shown). Further work is required to determine sex-specific ontological and morphological variation in volatile emission rate, site of production and phenological variation. The full list of volatile compounds that we found in the headspace analyses of *C. purpureus* is given in Supplementary Table 1.

Springtail preference assays. To determine whether springtails prefer one reproductive moss sex over the other, and then to test whether this was due to volatile cues, as suggested by the GC × GC–TOFMS analyses, we conducted two sets of preference assays with the moss *C. purpureus*. First, we used protocols modified from well-established springtail food preference assays to construct preference chambers from Petri dishes[26,27], and we conducted preference assays in these dishes comparing male and female intact (non-wounded) reproductive *C. purpureus* samples. For each assay, we used a Petri dish (55 × 15 mm), placed a 55-mm diameter piece of filter paper in the bottom of the dish and placed two smaller pieces of filter paper, separated by 1.5 cm, on top of the larger filter paper. The two comparison samples (5-mm diameter moss patches of intact shoots) were placed on the two smaller filter papers. Moss samples were from two Oregon populations (northeast Portland and Portland State University populations), and both males and females were producing gametoecia. Using a metal spatula, we placed 20 to 40 springtails in each dish between the moss samples, wrapped the dishes in parafilm, darkened them with foil and placed them in the growth chamber in which the springtails were reared. After 120 min, we removed the moss samples and filter paper, and determined the number of springtails within each moss sample, the number of springtails that did not occupy a moss sample and the number of moss shoots and moss reproductive structures per sample. Plants were dried in a drying oven at 60 °C for 48 h and the dry weight was determined. We conducted 24 assays with 491 springtails choosing specific moss samples. Springtails were never reused in assays. We found no significant difference in dry weight between male and female moss samples ($P = 0.95$; mean ± s.e.m. = 8.8 ± 1.2 mg and 9.1 ± 1.3 mg for males and females, respectively). However, male moss samples had significantly more shoots and gametoecia per shoot than did females ($P = 0.03$; mean ± s.e.m. = 25.7 ± 1.88 and 20.3 ± 1.6 for male and female shoots, respectively; $P = 0.001$; mean ± s.e.m. = 1.46 ± 0.32 and 0.22 ± 0.02, for gametoecia per male and female shoot, respectively).

To determine whether the preference that we found for female plants was due to springtails perceiving a volatile cue or another type of assessment (for example, visual), we set up a second set of assays using a static air olfactometer with intact (non-wounded) male and female *C. purpureus* samples. The olfactometer was a modified version of that described in a previous study[33], with the same additional modifications for springtails as described in another paper[28] The olfactometer was made of clear acrylic pipe with two sample compartments divided by a vertical plate. A walking arena for the springtail was placed above the compartments, with the springtails separated from the samples by a wetted opaque filter, to obscure visual choice. For each assay, a male and a female moss sample (15-mm diameter moss patches of intact shoots) were added to separate compartments of the olfactometer. Plants were collected in the field in May 2012 from several sites within a large population (>6,000 m^2) in North Plains, Oregon, and plants were nearing the end of the fertilization season, with females producing a few gametoecia and many new sporophytes, and males producing many gametoecia with ripe antheridia. Using a spatula, we placed 20 to 40 springtails on the walking arena, the olfactometer and placed it in the dark, and we recorded the springtails' choices every 30 min for 120 min. We conducted 10 assays with 276 springtails choosing specific moss samples. We dried and weighed the moss samples, as for the previous assays. We found no significant difference in dry weight or the number of shoots between male and female samples ($P = 0.99$; mean ± s.e.m. = 176.31 ± 18.86 mg and 160.00 ± 29.84 mg for the dry weight of males and females, respectively; $P = 0.53$; mean ± s.e.m. = 0.21 ± 0.03 and 0.21 ± 0.04 for male and female shoots, respectively). Male moss samples differed significantly from female samples in the number of gametoecia per shoot ($P = 0.0003$; mean ± s.e.m. = 0.43 ± 0.07 and 0.21 ± 0.04 for males and females, respectively).

Bryophyte microcosms. To determine the effect of springtails and water spray on sperm dispersal in mosses, we set up factorial experiments in which we manipulated springtail and water spray levels in *C. purpureus* and *B. argenteum*, and we counted sporophyte number as an estimate of fertilization success using a method described previously[34]. To establish microcosms, we propagated the moss on a substrate of a 2:1 mixture of propagation grade sand and peat moss. The mosses were propagated by chopping fresh plant material and distributing the chopped material evenly among microcosms (pots of 6.4 × 6.4 cm), with microcosms containing either *C. purpureus* or *B. argenteum*. For each moss species, microcosms contained plant material from a mix of three to five populations and were composed of both males and females. The microcosms were placed in seedling trays, watered from below and covered in humidomes, to create an enclosed habitat that was conducive to growth for both the springtails and mosses. The experiments were set up in Adaptis 1000 Conviron growth chambers (Pembina), enabling us to control for temperature, light and relative humidity (14 h light–10 h dark cycles with 18 °C light–8 °C dark; 150 μmol of photons $m^{-2} s^{-1}$; and 65% constant humidity). The microcosms were subjected to one of four treatments: springtails only, water spray only, springtails and water spray, neither springtails nor water spray. Microcosms of water spray and no-spray treatments were evenly distributed among trays of one of two designations (springtails or no springtails). One litre of water was maintained in the base of each tray, and each tray was covered with a humidome lid. Trays were rotated every 2 weeks within growth chambers to control for chamber effects. Water spray was applied approximately once per week with a squirt bottle containing room-temperature spring water. After 80 days in microcosms, an excess of algae accumulated in the *B. argenteum* microcosms, and the spray treatment was intermittent to allow the plants to recover; however, the spray was maintained at least every 14 days. Springtails (approximately 20 per microcosm) were added from stock cultures to all appropriate treatment trays once every 2 to 3 weeks and were observed living in the treatment microcosms in the weeks after application. We counted the number of sporophytes in each treatment after initial sporophyte formation (which we defined as the day when at least 15% of microcosms had sporophytes). The *B. argenteum* was started in August 2010 and took 44 days to reach initial sporophyte formation after planting. The *C. purpureus* were run as two separate experiments (starting in July 2010 and September 2010) with several trays per treatment for each set. One set took 231 days, whereas the second set took 179 days to reach initial sporophyte formation after planting.

We used a third experiment of *C. purpureus* ($n = 32$ microcosms) with the same treatments to test for variation among springtail treatments in the number of gametoecia, chlorophyll fluorescence of photosystem II efficiency (variable:maximum fluorescence (F_v/F_m)); and plant nitrogen content. We found no significant differences among springtail treatments in any of these measures, although adding springtails increased sporophyte production, as in the other experiments. These data suggest that the springtail addition did not enhance reproductive expression leading to more sporophyte production, and it did not alter overall plant health before sporophyte formation, consistent with a role of springtails in mediating sperm transfer.

Data analysis. Multivariate analysis was used to discriminate among volatile scent profiles[35]. Specifically, non-metric multidimensional scaling (NMDS) and analysis of similarities (ANOSIM) were carried out using the R Project for Statistical Computing to test for differences among volatile scent composition between male and female plants. Prior to analyses, individual volatile compounds were sorted into one of 21 IUPAC compounds classes and square-root transformed. All other analyses were conducted using JMP Version 10.0 (ref. 36). To test for differences in the average number of VOCs released between male and female plants, we used *t*-tests. To determine whether springtails chose preferentially between the two samples for each of the two types of preference assays (female versus male samples in Petri dishes or in the olfactometer) at 120 min, we used *G*-tests. We used springtail choice data from the olfactometer assays at 120 min only because there was no significant difference among time points. For preference assays, we used *t*-tests to determine whether male and female moss patch samples differed in dry weight, shoot number or number of gametoecia per shoot. We used logit analysis to determine the effect of the springtail treatment, the water spray treatment and the interaction between these treatments on the fraction of microcosms with sporophytes. We included seedling tray, nested in springtail treatment, in the model. For the logit analysis, we included species and interactions with species but found that these were not significant, and they were therefore dropped from the model. We also used a similar mixed-model nested analysis of variance (ANOVA) to analyse how the number of sporophytes per microcosm (log transformed) were affected by these factors.

29. Lawton, E. *Moss Flora of the Pacific Northwest.* (Hattori Botanical Laboratory, 1971).
30. Fountain, M. T. & Hopkin, S. P. *Folsomia candida* (Collembola): A ''standard'' soil arthropod. *Annu. Rev. Entomol.* **50,** 201–222 (2005).
31. Johnson, D. L. & Wellington, W. G. Predation of *Apochthonius minimus* (Pseudoscorpionida: Chthoniidae) on *Folsomia candida* (Collembola: Isotomidae) I. Predation rate and size-selection. *Res. Popul. Ecol. (Kyoto)* **22,** 339–352 (1980).
32. Xu, J., Ke, X., Krogh, P. H., Wang, Y., Lou, Y.-M. & Song, J. Evaluation of growth and reproduction as indicators of soil metal toxicity to the Collembolan, *Sinella curvis. Insect Sci.* **16,** 57–63 (2009).
33. Steidle, J. L. M. & Schöller, M. Olfactory host location and learning in the granary weevil parasitoid *Lariophagus distinguendus* (Hymenoptera: Pteromalidae). *J. Insect Behav.* **10,** 331–342 (1997).
34. Mishler, B. D. Reproductive biology and species distinctions in the moss genus *Tortula,* as represented in Mexico. *Syst. Bot.* **15,** 86–97 (1990).
35. van Dam, N. M. & Poppy, G. M. Why plant volatile analysis needs bioinformatics-detecting signals from noise in increasingly complex profiles. *Plant Biol.* **10,** 29–37 (2008).
36. SAS Institute. JMP for Windows. Release 10.0.0. (SAS Institute, 2012).

doi:10.1038/nature11497

Attention deficits without cortical neuronal deficits

Alexandre Zénon[1,2] & Richard J. Krauzlis[2,3]

The ability to process relevant stimuli selectively is a fundamental function of the primate visual system. The best-understood correlate of this function is the enhanced response of neurons in the visual cortex to attended stimuli[1,2]. However, recent results show that the superior colliculus (SC), a midbrain structure, also has a crucial role in visual attention[3–5]. It has been assumed that the SC acts through the same well-known mechanisms in the visual cortex[3,5]. Here we tested this hypothesis by transiently inactivating the SC during a motion-change-detection task and measuring responses in two visual cortical areas. We found that despite large deficits in visual attention, the enhanced responses of neurons in the visual cortex to attended stimuli were unchanged. These results show that the SC contributes to visual attention through mechanisms that are independent of the classic effects in the visual cortex, demonstrating that other processes must have key roles in visual attention.

Visual attention is a fundamental brain function that makes it possible to base perceptions and actions on the relevant parts of the environment. In the laboratory, visual attention is typically studied by asking subjects to respond to the properties of a cued stimulus while simultaneously ignoring the content of irrelevant, distracting stimuli. Twenty-five years ago, it was shown that in the primate visual cortex, the activity of neurons responsive to cued visual stimuli was higher than the activity evoked by un-cued distracters[6]. This finding, later termed 'gain modulation', has been subsequently observed in many different areas of the cerebral cortex[2,7,8], in many variants of the cueing task[8].

Visual attention is now understood to involve a network of areas, including the frontal and parietal cortex, as well as the visual cortex[9], and gain modulation of sensory responses is commonly considered to be the keystone of the neuronal mechanisms of attention[1,2].

Correlates of visual attention are not restricted to the cortex and have also been found in subcortical structures such as the SC[10,11] and thalamus[12–14]. Some of these effects could be inherited from the cortex. However, manipulation of neuronal activity in the SC alters or disrupts performance in tasks that test visual attention[3–5], indicating that the SC has a causal role. In a recent study using pharmacologic inactivation of the SC, monkeys had to report the direction of motion in a stimulus at a cued location, while ignoring equivalent motion in an irrelevant 'foil' stimulus located elsewhere[4]. After SC inactivation, the animals showed profound deficits in visual attention: they largely failed to report the direction of motion of the cued stimulus when it was placed in the part of the visual field affected by SC inactivation, and instead reported the direction of motion of the foil stimulus. Activity in the SC is therefore not simply updated about visual attention but seems to be necessary for its normal operation.

Previous studies have generally assumed that the SC plays a part in attention by influencing the well-known mechanisms in the visual cortex[3,5]. If so, then disrupting visual attention by inactivating the SC should change attention-related effects in the visual cortex. We tested this hypothesis by recording the activity of single neurons in the middle temporal area (MT) and medial superior temporal area (MST)—two cortical visual areas well known for their roles in processing motion signals[15] and their modulation by visual attention[16]—while monkeys performed a motion-change-detection task. We measured how neuronal activity was modulated by spatial cues before and during temporary pharmacological inactivation of SC. Contrary to the hypothesis, we found that attention-related effects in MT and MST remained intact even though SC inactivation caused major deficits in the visual attention task.

Two monkeys (J and M) performed a motion-detection task in which they were rewarded for pressing a button when they correctly detected a change in the direction of motion of the stimulus at the cued location and ignored changes in the direction of motion of a foil stimulus located diagonally opposite the cued stimulus (Fig. 1a). In trials in which the change occurred in the cued stimulus, the animals pressed the button correctly in about 50–60% of the trials (Fig. 1c, pre-injection 'hit rates' were 53 ± 26% for J and 57 ± 21% for M). Conversely, they correctly refrained from responding in most of the trials in which the change occurred in the distracter stimulus (pre-injection 'false alarm' rates were 9 ± 15% for J and 9 ± 7% for M).

To test the effects of SC inactivation on attention and sensory cortex activity during this task, we injected muscimol, a $\mathrm{GABA_A}$ (γ-aminobutyric acid type A) agonist, in the intermediate and deep layers of the SC (Fig. 1b). The extent of the neuronal inhibition caused by the injection was assessed at the beginning and end of each session, by measuring eye peak velocity during visually guided saccades[17]. Each session included two data-collection phases, one before and one during SC inactivation.

Consistent with previous results[4], we found that SC inactivation caused large and spatially specific deficits in the ability of the animal to detect changes in the cued stimulus, with post-injection hit rates dropping to about 10–15% in the part of the visual field affected by SC inactivation (Fig. 1 c–e and Supplementary Information). We then tested whether SC inactivation induced comparable changes in the cue-related modulation of activity in MT and MST.

Neurons were recorded in either the MT or MST area while the monkeys performed the task, during the same behavioural sessions documented above. The location and direction of motion of the stimuli were based on the tuning properties of the neurons, and the size of the motion patch was adjusted to the size of the receptive fields (see Methods). In brief, either the cued or the foil stimulus was placed in the receptive field of the neuron under study, and the direction of motion on each trial was set as the preferred or anti-preferred direction of the neuron, and was always opposite in the two stimulus patches. We recorded a total of 69 MST (monkey J, 31; monkey M, 38) and 44 MT (J, 34; M, 10) neurons before inactivation and 77 MST (J, 26; M, 51) and 55 MT (J, 47; M, 8) neurons during inactivation. Some of these neurons were isolated continuously throughout the experiment ($n = 36$ cells for MST and $n = 18$ cells for MT). We provide additional analyses for this particular set of neurons in Supplementary Information.

Before SC inactivation, as expected from previous studies demonstrating attention-related modulation of visual responses in MST and MT[16], we found that neurons recorded in MST (Fig. 2c) and MT (Fig. 2g) showed higher discharge rates when the motion stimulus in their receptive field was cued ('cue in') than when it was not cued ('cue out'). As in previous studies, we quantified this modulation by

[1]Institute of Neuroscience, Université Catholique de Louvain, 1200 Brussels, Belgium. [2]Systems Neurobiology Laboratory, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. [3]Laboratory of Sensorimotor Research, National Eye Institute, Bethesda, Maryland 20892, USA.

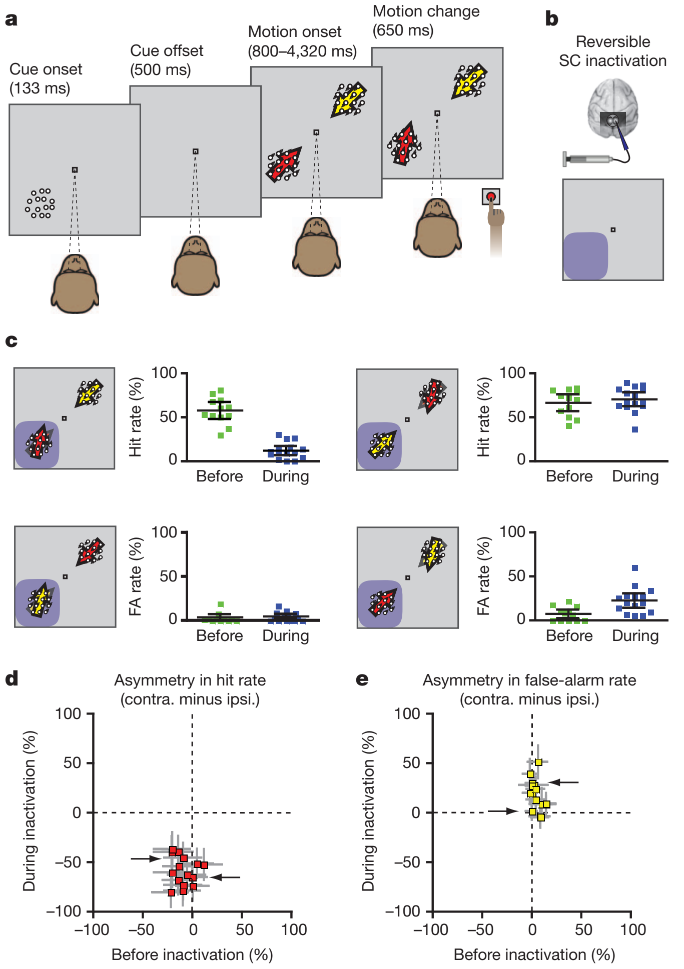


Figure 1 | Task design and behavioural performance. **a**, After a brief static cue, two motion stimuli moving in opposite directions were displayed in diagonally opposite locations. After a variable delay, the motion direction of one of the stimuli changed slightly. The monkey had to press a button when the change occurred at the cued location. **b**, We recorded single neurons in the MT or MST area after we injected muscimol into the intermediate and deep layers of the SC. The extent of the effect of the inactivation was assessed by mapping saccade velocities across the visual field. The affected part of the visual field is shown here schematically in blue. **c**, Response rates for changes at cued location (top) and un-cued location (bottom), before (green) and during (blue) SC inactivation. The red arrow denotes the cued patch and the yellow arrow denotes the un-cued patch. The affected part of the visual field is illustrated by blue shading. Error bars indicate 95% confidence intervals of the mean. FA, false alarm. **d**, **e**, Difference in cued change-detection rate (red) and false-alarm rate for un-cued motion changes (yellow) between the sides contralateral (contra.) and ipsilateral (ipsi.) to the injection before (*x* axis) and during (*y* axis) SC inactivation. Each dot corresponds to a different experiment and the grey lines show the 95% confidence interval (the computation of which is based on a method described in ref. 29). The arrows point to the data corresponding to the two sample experiments shown in Fig. 2.

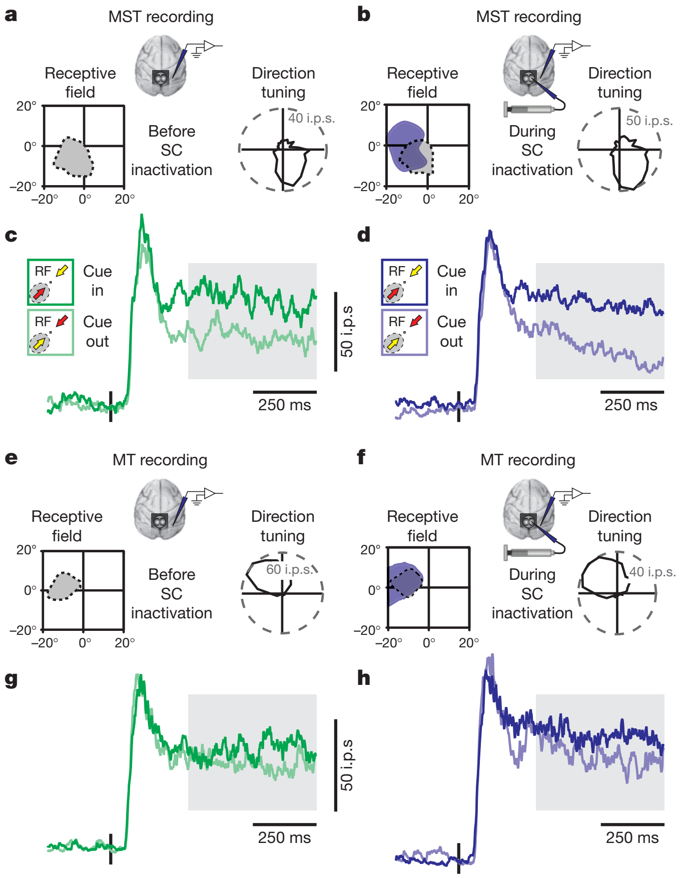


Figure 2 | Sample neuronal activity before and during SC inactivation. **a**, **b**, **e**, **f**, Receptive field and tuning properties of a sample MST (**a**, **b**) and MT (**e**, **f**) neuron recorded both before (**a**, **e**) and during (**b**, **f**) SC inactivation. The blue shading illustrates the extent of the effect of the muscimol injection in these experiments, on the basis of saccade velocities. **c**, **d**, **g**, **h**, Response of the same sample MST (**c**, **d**) and MT (**g**, **h**) neurons before (**c**, **g**, in green) and during (**d**, **h**, in blue) SC inactivation, for trials in which the cued patch was in (darker line) or out (lighter line) of the receptive field (RF). The vertical lines mark the onset of the motion stimuli. The grey box illustrates the time period used to compute the cue-related modulation analyses. ips, impulses per second.

measuring the discharge rate during the delay period of the task (300–800 ms after motion stimuli onset), and computed a modulation index, defined as the difference in discharge rates between cue in and cue out conditions, divided by their sum. For the two sample neurons shown in Fig. 2, the modulation indexes were 0.16 and 0.07 for the MST and MT neurons, which corresponded to increases in the discharge rate of 39% and 15%, respectively.

During SC inactivation, this modulation was intact. Neurons in MST (Fig. 2d) and MT (Fig. 2h) continued to show higher discharge rates for the motion stimulus in their receptive field when it was cued than when it was not cued. The post-injection modulation indexes were 0.21 and 0.08 for the MST neuron and MT neuron, respectively, which were not significantly different from their pre-injection values, but remained significantly greater than chance (both $P < 0.0001$, Wilcoxon rank-sum test, cue in versus cue out). This cue-related modulation in discharge rate was intact, despite the deficits in detection performance observed simultaneously during the SC inactivation (Fig. 1d).

To quantify the effect of SC inactivation across our population, we measured a modulation index for each neuron before and during inactivation. Pre-injection, the average modulation index in our sample of neurons was 0.075 ± 0.029 (mean ± 95% confidence interval; median, 0.051) in MST and 0.061 ± 0.023 in MT (median, 0.048; significantly greater than zero, Wilcoxon signed-rank test, all $P < 0.001$) (Fig. 3a) corresponding to average increases in the discharge rate of 24% and 15%, respectively. Post-injection, the average modulation index was 0.071 ± 0.025 (median, 0.048) in MST and 0.057 ± 0.022 in MT (median, 0.041) (Fig. 3a); these values remained significantly greater than zero (Wilcoxon signed-rank test, all $P < 0.001$), and were not different from the values before inactivation (Wilcoxon rank-sum test, $P > 0.5$; Bayesian posterior probability of the null (no-change) hypothesis ($p(H_0)$), MST, 0.99; MT, 0.985). Thus, SC inactivation produced no appreciable change in the cue-related modulation of the average discharge rate across our sample of MST and MT neurons. Similar

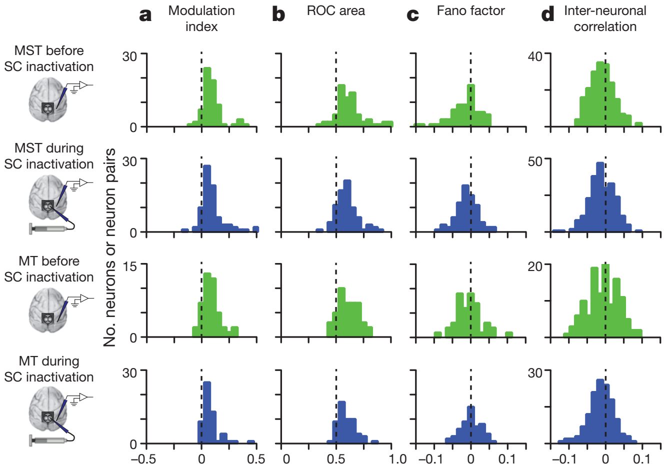


Figure 3 | Population results before and during SC inactivation. **a**–**d**, Distribution of modulation indices (**a**), area under ROC curves (**b**), Fano factor indices (**c**) and difference in interneuronal correlations (**d**) during the delay period, before (green) and during (blue) SC inactivation for all MST and MT neurons.

results were found when the non-preferred stimulus was presented inside the receptive field (Supplementary Information).

We considered whether SC inactivation might have altered other aspects of cue-related changes in MST and MT neuronal activity. Although modulation of average discharge rate is the standard method for documenting attention-related changes in neuronal activity, it does not measure how noise or variability of discharge rate might change with attention.

To address this point, we computed three additional values for each neuron. First, we computed the area under the receiver operating characteristic (ROC) curve[18], which indicates how well an ideal observer could classify the condition based on the activity of the neuron; in our case, whether the cued or un-cued stimulus was in the receptive field of the neuron. Second, we computed the Fano factor (the ratio of the variance over the mean of the response), which has been found to be lower for cued stimuli than for un-cued stimuli[19], indicating that attention decreases the variability of neuronal activity. Third, we computed the noise correlation between pairs of simultaneously recorded neurons, which has recently been found to decrease with attention[20,21], improving the signal-to-noise ratio of visual signals across the population of neurons.

These additional measurements were also unchanged by SC inactivation. The ROC areas were significantly higher than chance (Fig. 3b), both before and during SC inactivation in both MST and MT (Wilcoxon signed-rank test, all $P < 0.001$), and were not changed by SC inactivation (Wilcoxon rank-sum test, MST, $P = 0.30$; MT, $P = 0.28$; Bayesian $p(H_0)$, MST, 0.978; MT, 0.981); this result indicates that the ability of an ideal observer to discriminate the cued location was unchanged by SC inactivation.

The Fano factor index was significantly less than zero (Fig. 3c), both before (Wilcoxon signed-rank test, MST, $P < 0.0001$; MT, $P = 0.007$) and during (MST, $P < 0.0001$; MT, $P = 0.04$) inactivation, and not different from each other (MST, $P = 0.66$; MT, $P = 0.23$; Bayesian $p(H_0)$, MST, 0.988; MT, 0.979); this result shows that the variability in the discharge rate was reduced by spatial cueing both before and during SC inactivation.

The change in interneuronal correlation was significantly less than zero (Fig. 3d), both before (MST, $P < 0.0001$; MT, $P = 0.0001$) and during (MST, $P = 0.0001$; MT, $P = 0.0005$) inactivation, and not different from each other (MST, $P = 0.76$; MT, $P = 0.71$; Bayesian $p(H_0)$, MST, 0.996; MT, 0.995); this result indicates that spatial cues reduced the correlation in activity between neurons, and this reduction was unchanged by SC inactivation. Similar findings were made with a wide range of bin sizes used to compute the correlations (Supplementary Information).

Finally, we examined cue-related modulations in neuronal activity during other intervals in the task as well as changes in neuronal activity unrelated to the cue, and these were also unchanged during SC inactivation (Supplementary Information). We also confirmed that neuronal activity in the parts of MST we recorded were indeed necessary for the performance of the attention task (Supplementary Information).

In summary, we found that during SC inactivation, the enhanced responses of neurons in the visual cortex to attended stimuli were preserved despite large behavioural impairments in a covert attention task. This result was found in two visual areas well known for their roles in processing motion signals[15] and their modulation by visual attention[16]. Moreover, the attention deficit induced by SC inactivation not only preserved the cue-related changes in visual responses, but it also left intact the other known correlates of attention in the visual cortex: the ability of neurons to discriminate cued from un-cued spatial locations, the reliability of neuronal discharge (that is, Fano factor) and cue-related changes in noise correlations between neurons. These effects cannot be explained by a sensory impairment, because previous studies have shown that attention deficits during SC inactivation are not caused by changes in local motion perception[4]. The effects also cannot be explained by a motor deficit, because the single-button response in our task was unimpaired for stimuli outside the affected region of the visual field (Fig. 1c).

These findings demonstrate that the known modulations of activity in the visual cortex are not the only mechanisms involved in the control of attention and that other processes must have a key role. One possibility is that visual attention involves other aspects of neuronal activity in these same visual areas. For example, although we found no changes in correlations between nearby neurons, there could be changes between more distant sites or across different areas. A second possibility is that the crucial steps take place in other brain areas entirely, for example, in the parietal or prefrontal cortex[22], the SC or the basal ganglia[23]. In particular, the frontal eye fields (FEF) exert effects on attention qualitatively similar to the SC[24,25]. However, because of prominent feedback from FEF to the visual cortex, SC-induced changes in FEF might have been expected to also change responses in the visual cortex. Finally, it is possible that distinct circuits mediate different aspects of attention. For example, changes in the visual cortex might be important for feature-based attention[26] and for regulating the perceptual appearance of stimuli[27], whereas the mechanism targeted by SC inactivation is important for the all-or-none aspects of spatial attention (for example, change blindness[28]).

METHODS SUMMARY

We performed MT and MST neuronal recording and reversible inactivation of the SC in two adult rhesus monkeys (subjects J and M). The animals were prepared using standard surgical techniques described in detail in ref. 17. All experimental protocols were approved by the Institutional Animal Care and Use Committee and complied with US Public Health Service policy on the humane care and use of laboratory animals. The laboratory set-up for behavioural control and monitoring was identical to that described in ref. 17.

At the beginning of each inactivation session, we lowered a recording tetrode in a track selected on the basis of previous recording sessions. After identification of a good recording spot for MT and MST neurons, we mapped the receptive fields (see examples in Fig. 2a, b, e, f; 50–80 trials) and motion-direction tuning properties (Fig. 2a, b, e, f; 30–60 trials) of the isolated neurons and recorded them during performance of the attentional task (232–366 trials). After completion of the pre-injection data collection, an injectrode was lowered into the intermediate and deep layers of the SC and muscimol was injected following a procedure described in ref. 4. Around 20 min after the beginning of the injection, the extent of the effect of the inactivation was evaluated on the basis of eye velocity during visually guided saccades (60–120 trials). MT and MST neurons were then recorded again during the receptive field and tuning-mapping procedures and during the attentional task.

Full Methods and any associated references are available in the online version of the paper.

Received 30 January; accepted 9 August 2012.

Published online 12 September 2012.

1. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18,** 193–222 (1995).
2. Reynolds, J. H. & Chelazzi, L. Attentional modulation of visual processing. *Annu. Rev. Neurosci.* **27,** 611–647 (2004).
3. Müller, J. R., Philiastides, M. G. & Newsome, W. T. Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proc. Natl Acad. Sci. USA* **102,** 524–529 (2005).
4. Lovejoy, L. P. & Krauzlis, R. J. Inactivation of primate superior colliculus impairs covert selection of signals for perceptual judgments. *Nature Neurosci.* **13,** 261–266 (2010).
5. Cavanaugh, J. Subcortical modulation of attention counters change blindness. *J. Neurosci.* **24,** 11236–11243 (2004).
6. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229,** 782–784 (1985).
7. Treue, S. Neural correlates of attention in primate visual cortex. *Trends Neurosci.* **24,** 295–300 (2001).
8. Roelfsema, P. R., Lamme, V. A. & Spekreijse, H. Object-based attention in the primary visual cortex of the macaque monkey. *Nature* **395,** 376–381 (1998).
9. Corbetta, M. & Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nature Rev. Neurosci.* **3,** 201–215 (2002).
10. Kustov, A. A. & Robinson, D. L. Shared neural control of attentional shifts and eye movements. *Nature* **384,** 74–77 (1996).
11. Ignashchenkova, A., Dicke, P. W., Haarmeier, T. & Thier, P. Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention. *Nature Neurosci.* **7,** 56–64 (2003).
12. O'Connor, D. H., Fukui, M. M., Pinsk, M. A. & Kastner, S. Attention modulates responses in the human lateral geniculate nucleus. *Nature Neurosci.* **5,** 1203–1209 (2002).
13. Bender, D. B. & Youakim, M. Effect of attentive fixation in macaque thalamus and cortex. *J. Neurophysiol.* **85,** 219–234 (2001).
14. Robinson, D. L. & Petersen, S. E. The pulvinar and visual salience. *Trends Neurosci.* **15,** 127–132 (1992).
15. Rudolph, K. Transient and permanent deficits in motion perception after lesions of cortical areas MT and MST in the macaque monkey. *Cereb. Cortex* **9,** 90–100 (1999).
16. Treue, S. & Maunsell, J. H. R. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382,** 539–541 (1996).
17. Hafed, Z. M., Goffart, L. & Krauzlis, R. J. Superior colliculus inactivation causes stable offsets in eye position during tracking. *J. Neurosci.* **28,** 8124–8137 (2008).
18. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (1982).
19. Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron* **55,** 131–141 (2007).
20. Cohen, M. R. & Maunsell, J. H. R. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neurosci.* **12,** 1594–1600 (2009).
21. Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron* **63,** 879–888 (2009).
22. Moore, T. The neurobiology of visual attention: finding sources. *Curr. Opin. Neurobiol.* **16,** 159–165 (2006).
23. Redgrave, P., Prescott, T. J. & Gurney, K. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience* **89,** 1009–1023 (1999).
24. Moore, T. & Armstrong, K. M. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* **421,** 370–373 (2003).
25. Wardak, C., Ibos, G., Duhamel, J.-R. & Olivier, E. Contribution of the monkey frontal eye field to covert visual attention. *J. Neurosci.* **26,** 4228–4235 (2006).
26. Treue, S. & Martínez Trujillo, J. C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399,** 575–579 (1999).
27. Carrasco, M., Ling, S. & Read, S. Attention alters appearance. *Nature Neurosci.* **7,** 308–313 (2004).
28. Rensink, R. A., O'Regan, J. K. & Clark, J. J. To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci.* **8,** 368–373 (1997).
29. Ross, T. D. Accurate confidence intervals for binomial proportion and Poisson rate estimation. *Comput. Biol. Med.* **33,** 509–531 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Boehle, N. Dill and A. Karnik for technical assistance, and R. Wurtz for discussions and reading of the manuscript. This work was supported by the F.M. Kirby Foundation and the National Eye Institute Intramural Research Program at the National Institutes of Health.

Author Contributions A.Z. and R.J.K. designed and conducted the experiments and wrote the manuscript. A.Z. analysed the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.J.K. (richard.krauzlis@nih.gov).

METHODS

Monkey preparation. We performed MT and MST neuronal recordings and reversible inactivation of the intermediate and deep layers of the SC in two adult rhesus monkeys (subjects J and M) that were 12–16 years of age and weighed 14–16 kg. The monkeys were prepared using standard surgical techniques described in detail in ref. 17. All experimental protocols were approved by the Institutional Animal Care and Use Committee and complied with US Public Health Service policy on the humane care and use of laboratory animals. The laboratory set-up for behavioural control and monitoring was identical to that described in ref. 17.

Attentional task. Trials began with the appearance of a central dot on which the monkey had to fixate during the whole trial duration. Achievement of fixation triggered the display of a peripheral stimulus, the cue, consisting of a 5–7°-wide patch of static dots. The actual size of the patch was chosen as not to exceed the size of the receptive fields of the neurons being recorded. On each trial, the cue could be displayed at one of two possible locations, chosen randomly. One of these locations was chosen to be in the centre of the receptive fields of the recorded neurons and the other one was the symmetric location across the fixation point. The cue was displayed for 133 ms and was followed by a 500-ms delay, during which only the fixation point was displayed. Two patches of moving dots were then displayed at the two previously described locations. The dots were moving in opposite directions in the two patches, one of which being the preferred direction of the neurons being recorded. The characteristics of the stimulus have been described elsewhere[4]. In the present case, the dots had an eight-frame lifetime (corresponding to 107 ms). The direction of motion of each dot was drawn from a normal distribution centred on the direction of motion of the patch and with a 16° standard deviation.

The direction of motion of the patches remained constant for 800 ms plus a geometrically distributed delay of mean 480 ms (range, 0–3520 ms). This distribution allowed the hazard function to remain flat during the delay. After this delay, the direction of motion of one of the patches changed. The monkey had to press a button whenever the change in direction occurred at the previously cued location. The change varied from 16° to 20° and was adjusted on the basis of the performance of the monkey at the beginning of each session to keep a global performance of about 75%.

After the beginning of the change in direction, stimuli remained on the screen for 650 ms or until the response of the animal. Monkeys received a liquid reward only for correct responses in completed trials (button press after change occurred at cued location or absence of response when no change occurred or change occurred at un-cued location). If the monkey broke fixation midtrial, the trial was aborted and repeated later in the session. This paradigm has been referred to as a 'filtering' task because it requires the monkey to actively ignore stimulus changes at the un-cued location. The advantage of this task design is that correct performance requires the filtering out of signals from irrelevant distracter stimuli. This paradigm is similar to that used originally to demonstrate attentional modulation in areas MT and MST[19] and more recently to show a causal role of the SC in the control of spatial attention[4]; it is also similar to that described in ref. 30.

All stimuli were displayed on a cathode ray tube display with a refresh rate of 75 Hz. The background luminance of the monitor was 14 cd m^{-2}. Luminance of the fixation dot and of each dot in the patches was 50 cd m^{-2}. Subjects pushed buttons mounted on a button box at waist level within easy reach of the left hand. Each subject used only its left hand to push buttons.

Procedure. At the beginning of each inactivation session, we lowered a recording tetrode in a track selected on the basis of previous recording sessions. After identification of a good recording spot for MT and MST neurons, we mapped the receptive fields (see examples in Fig. 2a, b, e, f; 50–80 trials) and motion-direction tuning properties (Fig. 2a, b, e, f; 30–60 trials) of the isolated neurons and recorded them during performance of the attentional task (232–366 trials). After completion of the pre-injection data collection, an injectrode, the tip of which was previously sitting above the quadrigeminal cistern, was lowered into the intermediate and deep layers of the SC and muscimol was injected as per the procedure described in ref. 4. After the injection, the extent of the effect of inactivation was evaluated by measuring eye peak velocity during visually guided saccades (60–120 trials).

By carefully choosing the injection site on the basis of exploratory recordings, by adjusting the volume of muscimol injected (between 0.4 and 0.6 μl) and the orientation of the bevel of the injection cannula, we were able to localize the affected region of visual space such as to encompass in all experiments the contralateral visual-stimulus location used during the attentional task.

The affected part of the visual field was defined as the portion of space where velocities were inferior to the lower bound of the confidence interval ($\alpha = 0.05$) of the baseline velocities. Next, the MT and MST neurons were recorded again during the receptive field and tuning-mapping procedures and during the attentional task. The pre-injection part, including isolation and mapping of the MT and MST neurons, lasted for 1.5–2.5 h, the lowering of the injectrode and the injection lasted together about 4 min and the post-injection part lasted between 1 and 2 h. The duration of a whole session lasted between 3.5 and 4.5 h. There was a total of 12 successful sessions with SC inactivation combined with recordings before and during SC inactivation.

For the MST injection experiment (Supplementary Fig. 3), we first lowered the injectrode to a depth previously recognized as being 500 μm above the lower limit of MST. We then injected a first muscimol dose of 0.5 μl, moved the injectrode up 500 μm and injected again, and so forth up to the upper limit of the area. This led to a total of four injections.

Behavioural analysis. Performance in the task was evaluated by the number of correct and incorrect trials (binomial variable) in each condition (ipsilateral versus contralateral and before versus during inactivation). The statistical tests used to assess the significance of the behavioural change induced by the inactivations were logistic regressions, with each condition and their interaction used as categorical predictor variables. Inactivations were considered as having a significant effect when the P value for the interaction between the conditions ipsilateral versus contralateral and before versus during was inferior to 0.05. When conducting these analyses on all sessions together, subject identity was added as a random categorical predictor to take into account repeated measurements.

Neuronal recordings. Recordings were conducted with a tetrode (Thomas Recording GmBH). Neuronal signals were amplified, band-pass filtered and digitized (Plexon recording system). Neurons were isolated during the experiment to allow for online mapping of their receptive fields and motion-direction tuning properties. In parallel, all waveforms passing a manually set threshold were stored for offline sorting. Offline sorting was conducted first automatically (Klustakwik sorting algorithm[31]) and was then refined manually. On average, we recorded 7.5 neurons per experimental session.

For inactivation experiments, the four-channel waveforms and interspike interval distributions of each neuron isolated before muscimol injection was correlated with the waveforms and interspike interval distribution of each neuron isolated after injection[32]. We then used these correlation values to identify the neurons that were putatively the same before and during inactivation (see also Supplementary Information).

Motion-direction tuning and receptive-field mapping. After isolation of the neurons, the motion-direction tuning of the cells was first evaluated, following a procedure similar to that described in ref. 33. In brief, the monkey had to fixate on a central dot while a whole-screen patch of dots was moving coherently in a direction changing on every frame, leading to a circular motion. The direction of rotation (clockwise or anticlockwise) was selected randomly on every trial. The response of the isolated neurons as a function of the direction of motion of the patch was used to determine their preferred direction of motion.

After the direction-tuning procedure, the receptive fields of the neurons were assessed. The monkey had to fixate on a central dot while patches of dots moving coherently in the preferred direction of motion of the cells were displayed in quick succession at locations selected randomly from a grid encompassing the whole screen. Typically, 48 different locations were probed.

Bayesian analysis. In order to estimate the probability of an absence of difference between the pre- and post-injection data, we computed the Bayes factor for the comparison between a model assuming a change in mean value during inactivation and a model assuming no change (H_0). When necessary, data were transformed to achieve a normal distribution. We computed the Bayes factor by means of different methods: fractional Bayes factor[34], Bayesian information criterion[35] and Bayesian t-test based on the Savage–Dickey ratio test[36]. These different methods provided comparable results. We mention in the main text only the $p(H_0)$ computed with the fractional Bayes factor method.

Interneuronal correlations. Interneuronal correlations were computed following the same procedure as described in ref. 21. In brief, the delay period (between 300 and 800 ms following stimuli onset) was divided into non-overlapping bins (4, 6, 7, 11, 16, 22, 31, 45, 63, 83, 125, 250 or 500 ms long) in which spike counts were computed. The average spike count in each bin was subtracted out from the spike-count values to remove any stimulus-locked response variation. Similarly, the slow variation in discharge rate over consecutive trials was also removed by subtracting the Gauss-weighted smoothing of spike-count changes (σ = five trials). Pearson correlations were computed for all pairs of units having a minimum discharge rate of five spikes per second (MT before inactivation, 122 pairs; MT during, 134 pairs; MST before, 194 pairs; MST during, 235 pairs).

We then estimated the effect of attention on interneuronal correlations by computing the difference in correlations (cue in minus cue out) for MST and MT. These differences are shown in Supplementary Material for all bin sizes. To illustrate these results in the main article, we chose a bin size of 31 ms (shown in Fig. 3), on the basis of the timescale of interneuronal correlations estimated in ref. 37. Because the spike counts obtained with this bin size were not always normally distributed, we also performed the same analysis using non-parametric Spearman correlations and obtained similar results.

30. Palmer, J. & Moore, C. M. Using a filtering task to measure the spatial extent of selective attention. *Vision Res.* **49,** 1045–1064 (2009).
31. Harris, K. D., Henze, D. A., Csicsvari, J., Hirase, H. & Buzsáki, G. Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J. Neurophysiol.* **84,** 401–414 (2000).
32. Dickey, A. S., Suminski, A., Amit, Y. & Hatsopoulos, N. G. Single-unit stability using chronically implanted multielectrode arrays. *J. Neurophysiol.* **102,** 1331–1339 (2009).
33. Schoppmann, A. & Hoffmann, K. P. Continuous mapping of direction selectivity in the cat's visual cortex. *Neurosci. Lett.* **2,** 177–181 (1976).
34. Berger, J. & Pericchi, L. in *Model Selection* Vol 38 (ed. Lahiri, P) 135–207 (Institute of Mathematical Statistics Lecture Notes – Monograph Series, 2001).
35. Wagenmakers, E. J. A practical solution to the pervasive problems of *p* values. *Psychon. Bull. Rev.* **14,** 779–804 (2007).
36. Wetzelsls, R., Raaijmakers, J. G. W., Jakab, E. & Wagenmakers, E. J. How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian *t* test. *Psychon. Bull. Rev.* **16,** 752–760 (2009).
37. Bair, W., Zohary, E. & Newsome, W. T. Correlated firing in macaque visual area MT: time scales and relationship to behavior. *J. Neurosci.* **21,** 1676–1697 (2001).

doi:10.1038/nature11348

Lrp4 is a retrograde signal for presynaptic differentiation at neuromuscular synapses

Norihiro Yumoto[1], Natalie Kim[1] & Steven J. Burden[1]

Motor axons receive retrograde signals from skeletal muscle that are essential for the differentiation and stabilization of motor nerve terminals[1]. Identification of these retrograde signals has proved elusive, but their production by muscle depends on the receptor tyrosine kinase, MuSK (muscle, skeletal receptor tyrosine-protein kinase), and Lrp4 (low-density lipoprotein receptor (LDLR)-related protein 4), an LDLR family member that forms a complex with MuSK, binds neural agrin and stimulates MuSK kinase activity[2–5]. Here we show that Lrp4 also functions as a direct muscle-derived retrograde signal for early steps in presynaptic differentiation. We demonstrate that Lrp4 is necessary, independent of MuSK activation, for presynaptic differentiation *in vivo*, and we show that Lrp4 binds to motor axons and induces clustering of synaptic-vesicle and active-zone proteins. Thus, Lrp4 acts bidirectionally and coordinates synapse formation by binding agrin, activating MuSK and stimulating postsynaptic differentiation, and functioning in turn as a muscle-derived retrograde signal that is necessary and sufficient for presynaptic differentiation.

Postsynaptic muscle cells provide signals to motor axons, and these signals regulate the formation, maturation, stabilization and plasticity of neuromuscular synapses[1]. During development, motor axons approach and form synapses with muscle in a prepatterned region, marked by elevated expression and clustering of key postsynaptic proteins, including acetylcholine receptors (AChRs)[6–11]. Muscle prepatterning depends on MuSK and Lrp4, which forms a complex with MuSK and stimulates MuSK kinase activity[3–7,9,12–14]. Stabilization of developing synapses requires motor-neuron-derived agrin, which binds Lrp4, stimulates further association between Lrp4 and MuSK, and increases MuSK kinase activity, leading to anchoring of key proteins in the postsynaptic membrane and elevated transcription of 'synaptic genes' in myofibre synaptic nuclei[4,5,13,15–17].

Lrp4 and MuSK are both required for presynaptic as well as postsynaptic differentiation, as motor axons grow beyond the prepatterned region and fail to cluster synaptic vesicles in mice deficient in either gene[2,3]. How agrin, Lrp4 and MuSK control presynaptic differentiation is poorly understood. Because Lrp4 activates MuSK, the presynaptic defects in *Lrp4*-mutant mice could be a consequence of inadequate MuSK activation and a failure to produce novel retrograde signals. Alternatively, Lrp4 may have a direct role in regulating motor axon growth and differentiation. To distinguish between these possibilities, we established a cell culture assay to determine whether Lrp4 is sufficient to induce presynaptic differentiation. First, we co-cultured motor neurons, dissected from *HB9*::*GFP* (green fluorescent protein) transgenic mice, with skeletal muscle cells and established culture conditions that were permissive for presynaptic differentiation. Under these conditions, synapsin, a protein that is associated with synaptic vesicles, accumulated in motor axons at sites that were apposed to AChR clusters in muscle (Fig. 1a, b). We then co-cultured motor neurons with NIH 3T3 cells or 3T3 cells (mouse embryonic fibroblast cell lines) expressing Lrp4 and stained for synapsin. Figure 1 shows that synapsin is distributed homogenously in axons of motor neurons co-cultured with control 3T3 cells, whereas synapsin accumulated in motor axons at sites of contact with Lrp4-expressing 3T3 cells (Fig. 1c). We also co-cultured motor neurons with HEK 293 cells that expressed a Flag-tagged version of Lrp4, allowing us to visualize cell surface Lrp4. Lrp4 was clustered on the cell surface, and synapsin accumulation in motor axons was often apposed to these clusters of Lrp4 (Figs 1d and Supplementary Fig. 2; see below). In addition, we transfected 293 cells with truncated forms of Lrp4 and found that the LDLa repeats from the extracellular region of Lrp4, in the absence of the EGF-like and β-propeller domains, are sufficient to induce presynaptic differentiation (Supplementary Fig. 2).

These experiments indicated that Lrp4 is sufficient to trigger presynaptic differentiation but left open the possibility that Lrp4 acted together with other proteins expressed in 3T3 and 293 cells to induce presynaptic differentiation. Therefore, we treated *HB9*::*GFP* motor neurons with an Lrp4–Fc fusion protein that contained the LDLa repeats and was attached to polystyrene microspheres, and stained for synapsin. We found that synapsin, as well as synaptophysin and SV2, bona-fide synaptic vesicle proteins, were clustered at contact sites with Lrp4-LDLa–Fc beads (Figs 2a and Supplementary Fig. 3). We also stained for bassoon, a protein that is concentrated at synaptic vesicle fusion sites in nerve terminals, called active zones, and found that bassoon was similarly clustered with Lrp4-LDLa–Fc beads (Fig. 2b). In contrast, neither Fc alone nor the LDLa repeats of Lrp1, another Lrp-family member, induced clustering of synapsin or bassoon, indicating that presynaptic differentiation is induced selectively by the LDLa repeats from Lrp4 (Figs 2b, c and Supplementary Fig. 3). Moreover, addition of soluble, dimeric Lrp4-LDLa–Fc, unattached to beads, failed to induce presynaptic differentiation (Supplementary Fig. 4), indicating that a large number of interactions, conferred by the attachment of the extracellular region of Lrp4 (ecto-Lrp4) to polystyrene microspheres, cooperate to mediate presynaptic differentiation.

Because MuSK, like Lrp4, is required for presynaptic differentiation *in vivo*, and because MuSK activation causes clustering of MuSK as well as Lrp4 at synapses, we tested whether the extracellular region of MuSK could also induce presynaptic differentiation. Although microspheres with ecto-MuSK–Fc or ecto-Lrp4–Fc attached equally well to motor axons (Supplementary Fig. 5), only ecto-Lrp4–Fc induced clustering of synaptic-vesicle and active-zone proteins (Fig. 2b). Moreover, Myc-tagged MuSK, expressed in 293 cells, failed to induce clustering of synapsin (Supplementary Fig. 2). Thus, although Lrp4 and MuSK are both required for the differentiation of motor nerve terminals *in vivo*, only Lrp4 is sufficient to stimulate presynaptic differentiation.

Because Lrp4 binds neural agrin, we asked whether agrin was required for Lrp4 to induce presynaptic differentiation. We crossed agrin-null ($Agrn^{-/-}$) and *HB9*::*GFP* mice and treated $Agrn^{-/-}$ explants with Lrp4-LDLa–Fc. Wild-type and $Agrn^{-/-}$ motor neurons were equally responsive to Lrp4-LDLa–Fc beads (Fig. 2d), indicating that Lrp4 induces presynaptic differentiation in a manner that does not depend on agrin.

[1] Molecular Neurobiology Program, Helen L. and Martin S. Kimmel Center for Biology and Medicine at the Skirball Institute of Biomolecular Medicine, New York University Medical School, 540 First Avenue, New York, New York, USA.

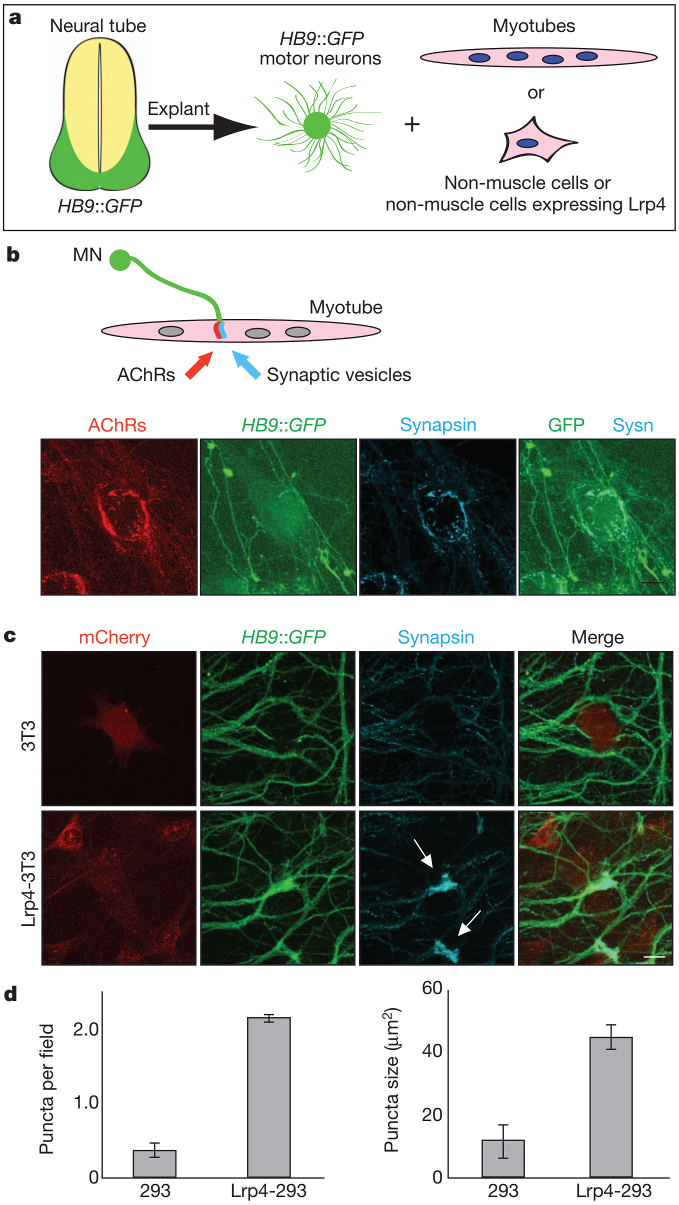


Figure 1 | Lrp4-expressing non-muscle cells induce clustering of synapsin in motor axons. a, Explants from the ventral neural tube of *HB9::GFP* transgenic mice, containing GFP-expressing motor neurons, were co-cultured with primary muscle cells or non-muscle cells. **b**, Synapsin (Sysn; blue) accumulates in motor axons (green) in apposition to clusters of AChRs (red) that form in muscle, marking synaptic sites. MN, motor neuron. **c**, Synapsin is homogenously distributed in motor neurons that are co-cultured with control 3T3 cells, expressing mCherry alone, but it is clustered (arrows) in motor axons that contact 3T3 cells expressing Flag–Lrp4–mCherry (Lrp4-3T3) **d**, The number and size of synapsin puncta are fivefold greater in axons contacting 293 cells expressing Flag–Lrp4 than control cells (mean ± s.e.m., $n = 3$). Scale bar, 10 µm.

To determine whether Lrp4 induced functional release sites, we measured recycling of synaptic vesicles using the styryl dye FM 4-64FX. Depolarization of motor neurons treated with Lrp4-LDLa–Fc beads, caused uptake of FM 4-64FX, and further depolarization led to the release of dye (Fig. 2e, f). In contrast, depolarization of motor neurons treated with Lrp1-LDLa–Fc beads, led to low and uniform axonal uptake of FM 4-64FX (Fig. 2e). Thus, Lrp4 induced both morphologically and functionally specialized neurotransmitter release sites.

Our experiments indicate that Lrp4 interacts with a protein expressed by motor axons to promote presynaptic differentiation. To determine whether motor axons express an Lrp4-binding protein,

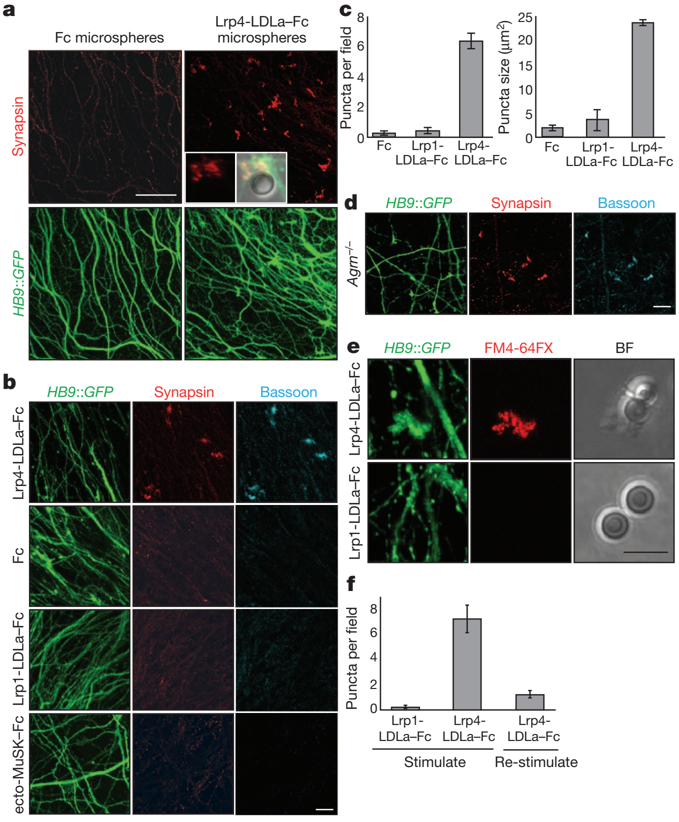


Figure 2 | Lrp4, attached to polystyrene beads, induces presynaptic differentiation in motor neurons. a, Lrp4-LDLa–Fc, attached to polystyrene beads, induces clustering of synapsin (red) in motor axons (green). Many synapsin clusters are in close apposition with Lrp4 beads (insets). Scale bar, 50 µm. **b**, Lrp4 specifically induces clustering of bassoon (blue) as well as synapsin (red). Scale bar, 10 µm. **c**, Lrp4 beads induce an approximately12-fold increase in the number of synapsin puncta (mean ± s.e.m., $n = 3$). **d**, Lrp4 beads induce synaptic puncta, marked by synapsin and bassoon, in *Agrn*-mutant motor neurons. The response of *Agrn*-mutant motor neurons is not significantly different from wild-type motor neurons (the response of *Agrn*-mutant motor neurons is 106% that of wild-type motor neurons). Scale bar, 10 µm. **e**, **f**, Depolarization stimulates uptake (Stimulate) and release (Restimulate) of FM 4-64FX in motor axons at contact sites with Lrp4-LDLa beads, visualized by bright field (BF) microscopy (mean ± s.e.m., $n = 3$). The cartoon shows the LDLa repeats and β-propeller domains in Lrp4. Scale bar, 5 µm.

we cultured explants from the ventral neural tube, which contains motor neurons, and probed the explants with an alkaline phosphatase (AP)–ecto–Lrp4 fusion protein. We stained for AP activity and found that AP–ecto-Lrp4 bound strongly to motor axons and preferentially along distal rather than proximal segments (Fig. 3a, b) indicating that motor neurons express an Lrp4-binding protein that is enriched approximately 30-fold on distal motor axons (Supplementary Fig. 6). The gradual and linear increase in binding from proximal to distal regions is probably due to an increase in number rather than affinity of Lrp4-binding sites, as preferential binding to distal segments is evident at the highest concentration (25 nM) of AP–ecto-Lrp4 that we tested (Supplementary Fig. 6). Binding of ecto-Lrp4 to motor axons is independent of agrin and mediated by the LDLa repeats from Lrp4 (Fig. 3b, c), mirroring the manner in which Lrp4 induces clustering of synaptic-vesicle and active-zone proteins. AP–ecto-Lrp4 also bound to axons emanating from dorsal neural tube explants, which lack motor neurons, although staining was less intense and more uniform compared to motor axons (Supplementary Fig. 7).

We next sought to determine whether Lrp4 is essential for motor axons to terminate and differentiate *in vivo*. Previously, we showed that increasing *Musk* expression in muscle of *Agrn*-mutant mice is

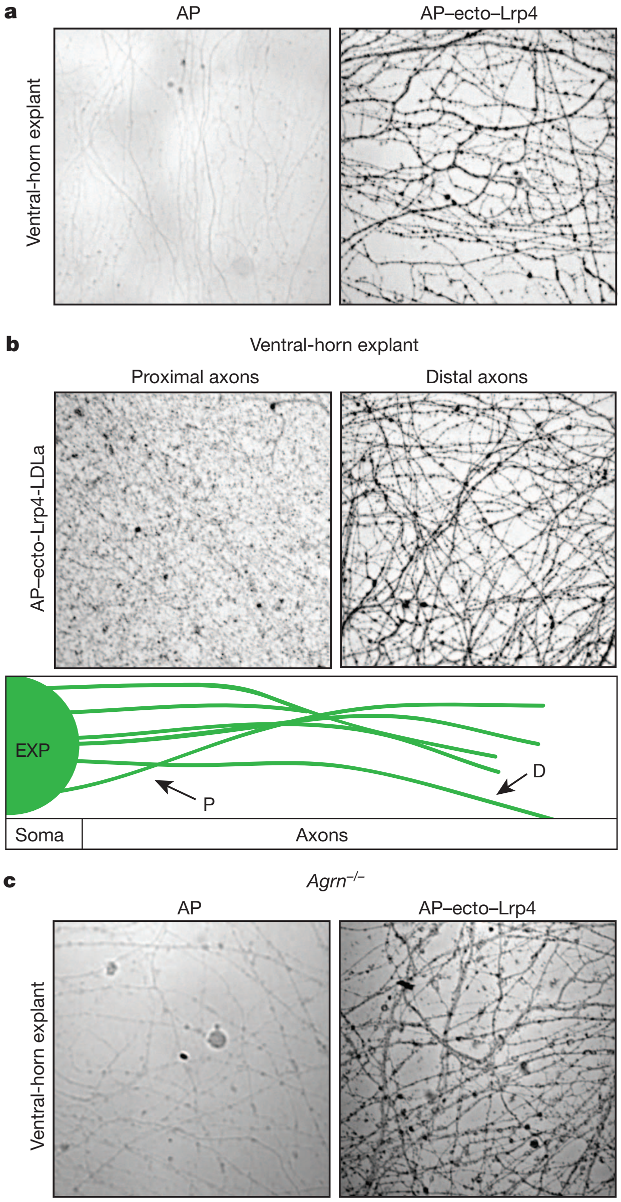


Figure 3 | Lrp4 binds to motor axons. **a**, AP–ecto-Lrp4 binds to motor axons. **b**, AP–ecto-Lrp4-LDLa binds preferentially to distal (D) rather than proximal (P) segments of motor axons extending from ventral horn explants (EXP). **c**, AP–ecto-Lrp4 binds to distal segments of agrin-mutant motor axons.

sufficient to rescue AChR clustering and presynaptic differentiation, preventing the neonatal lethality of *Agrn*-mutant mice[18]. These experiments showed that a modest increase in *Musk* expression can bypass the requirement for agrin and indicated that agrin normally acts to ensure that there is sufficient MuSK kinase activity to stabilize presynaptic and postsynaptic differentiation.

To determine whether *Musk* overexpression could bypass the requirement for Lrp4 in synapse formation, we crossed transgenic mice, which carry a human skeletal actin (*HSA*)-*Musk-L* transgene and express threefold more *Musk* in muscle than wild-type mice, with *Lrp4*-mutant mice and analysed diaphragm muscles from mice at embryonic day 18.5 (E18.5). In the absence of Lrp4, AChRs fail to cluster, and motor axons grow without terminating or differentiating (Fig. 4)[3]. *Musk* overexpression fully restored AChR clustering in *Lrp4*-mutant mice (Figs 4 and Supplementary Fig. 8), indicating that *Musk* overexpression can bypass the normal requirement for Lrp4 in postsynaptic differentiation. However, *Musk* overexpression failed to rescue presynaptic differentiation in *Lrp4*-mutant mice. Instead, motor axons grew throughout the muscle and rarely contacted AChR clusters (Fig. 4 and Supplementary Fig. 9). Moreover, *Musk* overexpression did not rescue the neonatal lethality of *Lrp4*-mutant mice, which is caused by a failure to form neuromuscular synapses[19]. These findings show that Lrp4 has an essential and early role, independent of MuSK activation, in presynaptic differentiation *in vivo*, as Lrp4 is required to arrest motor axon growth and induce clustering of synaptic vesicles.

We have a good, although incomplete, understanding of the signals and mechanisms for postsynaptic differentiation at neuromuscular synapses, and this knowledge has led to the identification of genes responsible for congenital myasthenia and the synaptic proteins that are targeted in autoimmune myasthenia gravis[20,21]. In contrast, discovery of the signals and mechanisms by which muscle cells control the differentiation of motor nerve terminals has proved more challenging and remains one of the notable gaps in our understanding of neuromuscular synapses.

Here we show that Lrp4 acts in a bidirectional manner, coordinating synaptic development, as Lrp4 not only binds agrin and regulates postsynaptic differentiation but also functions as a muscle-derived retrograde signal for early steps in presynaptic differentiation. This dual role of Lrp4 in presynaptic and postsynaptic differentiation represents a parsimonious means for mediating reciprocal signalling between adjacent cells and resembles the dual roles that Eph receptors and ErbB receptors have in responding to their respective ligands and in stimulating signalling in ligand-presenting cells[22,23]. Our findings suggest that Lrp4 functions as a critical check-point at three steps during synapse formation (Supplementary Fig. 1): first, before innervation, Lrp4 forms a complex with MuSK to establish muscle prepatterning; second, as motor axons approach muscle, Lrp4, clustered as a consequence of MuSK activation, acts as a retrograde signal to promote their differentiation; and third, once motor axons establish contact with muscle, Lrp4 binds agrin, which is released from motor nerve terminals, stimulating further MuSK phosphorylation and stabilizing neuromuscular synapses.

Other ligands, including members of the FGF7, FGF10 and FGF22 family, laminin β2, collagen IV and SIRPα, stimulate clustering of synaptic vesicles in cultured motor neurons and have a role in synaptic maturation *in vivo*[24]. Nevertheless, motor axons terminate and differentiate to a considerable extent in the absence of these signalling components, indicating that additional retrograde organizers regulate earlier steps in presynaptic differentiation[24]. Because motor axons fail to stop and display any signs of presynaptic differentiation in mice lacking Lrp4, Lrp4 must act at an early stage in presynaptic differentiation.

Auto-antibodies to AChRs, MuSK or Lrp4 are responsible for myasthenia gravis[25]. The clinical and pathological manifestations of anti-Lrp4 myasthenia have not been described in detail, but our studies indicate that auto-antibodies to Lrp4 have the potential to obstruct synaptic function not only by blocking binding between Lrp4 and agrin, or Lrp4 and MuSK, but also by interfering with binding between Lrp4 and Lrp4 receptors on nerve terminals. Because the premature withdrawal of motor nerve terminals, which causes muscle denervation, is an early step in amyotrophic lateral sclerosis and a characteristic feature of muscle wasting during ageing[26,27], defects in retrograde signalling may underlie or contribute to neuromuscular diseases and sarcopenia.

Lrp4, like *Musk*, is expressed in the cerebellum, cortex, hippocampus and olfactory bulb (see http://www.brainatlas.org), raising the possibility that Lrp4 may regulate synaptic differentiation in the central nervous system (CNS). Although *Lrp4*-mutant mice die at birth[3], well before the peak period of synapse formation in the CNS, *Lrp4*-mutant mice rescued for Lrp4 expression in muscle survive as adults, and should provide a good model system for studying the role of Lrp4 in synapse formation in the CNS[19].

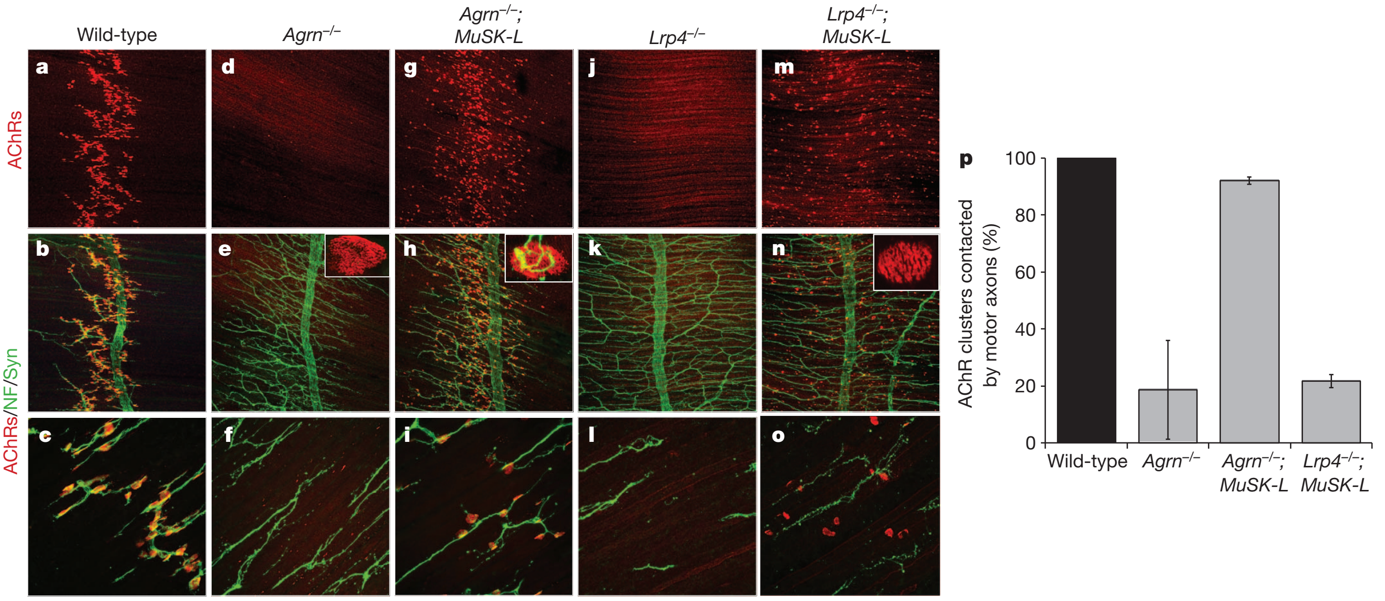


Figure 4 | Lrp4 is essential for presynaptic differentiation independent of MuSK activation. a–i, In *Agrn*-mutant mice, a threefold increase in *Musk* expression, conferred by the *Musk-L* transgene, restores AChR clusters and presynaptic differentiation[18]. **j–o**, In *Lrp4*-mutant mice, *Musk-L* restores AChR clusters but not nerve-terminal differentiation; instead, motor axons continue to grow beyond the prepatterned zone and fail to contact AChR clusters. **p**, In *Agrn*-mutant mice carrying *Musk-L*, approximately 90% of AChR clusters are contacted by motor axons. In *Lrp4*-mutant mice that carry *Musk-L*, approximately 15% of AChR clusters are contacted by motor axons; these contacts may be incidental, as motor axons grow and branch extensively throughout muscle of *Lrp4*-mutant mice, inevitably placing axons in the vicinity of AChR clusters (mean ± s.e.m., $n = 3$). The insets show higher-magnification views of AChR clusters that are innervated by motor axons (**h**) or devoid of contact from motor axons (**e**, **n**). NF, neurofilament; Syn, synaptophysin.

METHODS SUMMARY

Muscles from wild-type, agrin-mutant and *Lrp4*-mutant mice were stained with antibodies to neurofilament and synaptophysin to assess presynaptic differentiation, and with α-bungarotoxin to measure postsynaptic differentiation. Explants of neural tube, containing motor neurons, were grown in cell culture together with muscle, control non-muscle cells or non-muscle cells expressing Lrp4 or MuSK. Alternatively, motor neurons were treated with polystyrene microspheres, which had the extracellular region of Lrp4, Lrp1 or MuSK attached to the beads. Presynaptic differentiation was measured by staining with antibodies to presynaptic proteins and by quantifying vesicle recycling with the styryl dye FM4-64FX. Binding of AP–ecto-Lrp4 to the cell surface of motor axons was visualized and quantitated by staining for alkaline phosphatase activity.

Full Methods and any associated references are available in the online version of the paper.

Received 20 January; accepted 22 June 2012.
Published online 1 August 2012.

1. Sanes, J. R. & Lichtman, J. W. Induction, assembly, maturation and maintenance of a postsynaptic apparatus. *Nature Rev. Neurosci.* **2,** 791–805 (2001).
2. DeChiara, T. M. *et al.* The receptor tyrosine kinase MuSK is required for neuromuscular junction formation *in vivo*. *Cell* **85,** 501–512 (1996).
3. Weatherbee, S. D., Anderson, K. V. & Niswander, L. A. LDL-receptor-related protein 4 is crucial for formation of the neuromuscular junction. *Development* **133,** 4993–5000 (2006).
4. Kim, N. *et al.* Lrp4 is a receptor for Agrin and forms a complex with MuSK. *Cell* **135,** 334–342 (2008).
5. Zhang, B. *et al.* LRP4 serves as a coreceptor of agrin. *Neuron* **60,** 285–297 (2008).
6. Arber, S., Burden, S. J. & Harris, A. J. Patterning of skeletal muscle. *Curr. Opin. Neurobiol.* **12,** 100–103 (2002).
7. Yang, X. *et al.* Patterning of muscle acetylcholine receptor gene expression in the absence of motor innervation. *Neuron* **30,** 399–410 (2001).
8. Yang, X., Li, W., Prescott, E. D., Burden, S. J. & Wang, J. C. DNA topoisomerase IIbeta and neural development. *Science* **287,** 131–134 (2000).
9. Lin, W. *et al.* Distinct roles of nerve and muscle in postsynaptic differentiation of the neuromuscular synapse. *Nature* **410,** 1057–1064 (2001).
10. Panzer, J. A. S. o. n. g. Y. & Balice-Gordon, R. J. *In vivo* imaging of preferential motor axon outgrowth to and synaptogenesis at prepatterned acetylcholine receptor clusters in embryonic zebrafish skeletal muscle. *J. Neurosci.* **26,** 934–947 (2006).
11. Flanagan-Steet, H., Fox, M. A., Meyer, D. & Sanes, J. R. Neuromuscular synapses can form *in vivo* by incorporation of initially aneural postsynaptic specializations. *Development* **132,** 4471–4481 (2005).
12. Burden, S. J. SnapShot: neuromuscular junction. *Cell* **144,** 826.e1 (2011).
13. Kummer, T. T., Misgeld, T. & Sanes, J. R. Assembly of the postsynaptic membrane at the neuromuscular junction: paradigm lost. *Curr. Opin. Neurobiol.* **16,** 74–82 (2006).
14. Zhang, W., Coldefy, A. S., Hubbard, S. R. & Burden, S. J. Agrin binds to the N-terminal region of Lrp4 and stimulates association between Lrp4 and the first Ig-like domain in MuSK. *J. Biol. Chem.*, (2011).
15. Gautam, M. *et al.* Defective neuromuscular synaptogenesis in agrin-deficient mutant mice. *Cell* **85,** 525–535 (1996).
16. Lin, W. *et al.* Neurotransmitter acetylcholine negatively regulates neuromuscular synapse formation by a Cdk5-dependent mechanism. *Neuron* **46,** 569–579 (2005).
17. Misgeld, T., Kummer, T. T., Lichtman, J. W. & Sanes, J. R. Agrin promotes synaptic differentiation by counteracting an inhibitory effect of neurotransmitter. *Proc. Natl Acad. Sci. USA* **102,** 11088–11093 (2005).
18. Kim, N. & Burden, S. J. MuSK controls where motor axons grow and form synapses. *Nature Neurosci.* **11,** 19–27 (2008).
19. Gomez, A. M. & B. u. r. d. e. n. S. J. The extracellular region of Lrp4 is sufficient to mediate neuromuscular synapse formation. *Dev. Dynam.* **240,** 2626–2633 (2011).
20. Engel, A. G., Ohno, K. & Sine, S. M. Sleuthing molecular targets for neurological diseases at the neuromuscular junction. *Nature Rev. Neurosci.* **4,** 339–352 (2003).
21. Higuchi, O., Hamuro, J., Motomura, M. & Yamanashi, Y. Autoantibodies to low-density lipoprotein receptor-related protein 4 in myasthenia gravis. *Ann. Neurol.* **69,** 418–422 (2011).
22. Drescher, U. The Eph family in the patterning of neural development. *Curr. Biol.* **7,** R799–R807 (1997).
23. Bao, J., Wolpowitz, D., Role, L. W. & Talmage, D. A. Back signaling by the Nrg-1 intracellular domain. *J. Cell Biol.* **161,** 1133–1141 (2003).
24. Fox, M. A. *et al.* Distinct target-derived signals organize formation, maturation, and maintenance of motor nerve terminals. *Cell* **129,** 179–193 (2007).
25. Higuchi, O., Hamuro, J., Motomura, M. & Yamanashi, Y. Autoantibodies to low-density lipoprotein receptor-related protein 4 in myasthenia gravis. *Ann. Neurol.* **69,** 418–422 (2011).
26. Pun, S., Santos, A. F., Saxena, S., Xu, L. & Caroni, P. Selective vulnerability and pruning of phasic motoneuron axons in motoneuron disease alleviated by CNTF. *Nature Neurosci.* **9,** 408–419 (2006).
27. Valdez, G. *et al.* Attenuation of age-related changes in mouse neuromuscular synapses by caloric restriction and exercise. *Proc. Natl Acad. Sci. USA* **107,** 14863–14868 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Landmesser and K. Hata for helping us to establish a motor-neuron and muscle co-culture system, W. Zhang for constructs encoding Fc- and AP-tagged forms of Lrp4, T. Jessell and J. Dasen for *HB9::GFP* mice, J. Sanes for agrin-mutant mice, K. Anderson and L. Niswander for *Lrp4*-mutant mice, and D. Littman and R. Lehmann for comments on the manuscript. We are grateful to P. Lopez and M. J. Sunshine and personnel in the transgenic mouse core for their assistance. The flow cytometry and transgenic mouse cores are supported by a New York University Cancer Institute Center Support Grant (National Institutes of Health (NIH)/National Cancer Institute (NCI) 5 P30CA16087-31). This work was supported with funds from the National Institutes of Health (NS36193 to S.J.B.), the Skirball Institute and postdoctoral training support to N.Y. from New York State Stem Cell Science (NYSTEM).

Author Contributions N.Y. designed and carried out all of the experiments in Figs 1, 2 and 3. N.K. designed and carried out the experiments in Fig. 4. S.J.B. helped to design and interpret experiments. All authors wrote and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.J.B. (burden@saturn.med.nyu.edu).

METHODS

Co-culture of motor neurons and muscle or non-muscle cells. Explants of neural tube from *HB9*::*GFP* transgenic mice at embryonic day 11.5 (E11.5)–E13 were dissected and cultured in Neurobasal medium (Invitrogen), supplemented with B27 and GlutaMax (Invitrogen), 2 ng ml^{-1} BDNF, 2 ng ml^{-1} GDNF (Cell Sciences), 2 ng ml^{-1} CTNF, 1 ng ml^{-1} NGF (Sigma) and antibiotics. The ventrolateral portion of the neural tube, containing motor neurons, was dissected and isolated based on *HB9*::*GFP* expression; the dissected dorsal region of the neural tube lacked GFP expression. Explants were cultured on poly-L-ornithine- and laminin-coated tissue culture dishes for 4 to 6 days before application of microspheres or addition of myotubes or non-muscle cells (NIH 3T3 or HEK 293 cells). Mouse myotubes were generated from primary myoblasts in a separate culture dish and transferred to explant cultures by non-enzymatically detaching myotubes, as described previously[28]. Non-muscle cells were transfected with Flag–Lrp4, Flag–Lrp4–mCherry or mCherry[4], sorted by flow cytometry for mCherry or cell-surface Flag expression, using M2 antibodies (Sigma). We monitored Lrp4 expression either by viewing mCherry expression in cells transfected with Flag–Lrp4–mCherry (Fig. 1c) or by staining for Flag in cells transfected with Flag–Lrp4 (Figs 1d and Supplementary Fig. 1). Non-muscle cells were co-cultured with explants for 20 to 24 h in supplemented Neurobasal medium together with conditioned medium from rat Schwann cells or E12.5 mouse neural tube cells. Half of the medium was replaced every other day.

Assays for presynaptic differentiation. Co-cultures were fixed with 3.7% formaldehyde and stained with antibodies to synapsin (Synaptic Systems), GFP (Abcam), bassoon (Stressgen), SV2 (Developmental Studies Hybridoma Bank), Synaptophysin (Invitrogen) and Alexa 647-conjugated α-bungarotoxin (α-BGT; Invitrogen). Human Fc (Jackson ImmunoResearch), Lrp1-LDLa–Fc (Cluster II of Lrp1 from R&D Systems) or Lrp4-LDLa–Fc[14] were attached to Protein A microspheres (Bangslabs) and incubated with explants for 20 to 24 h in the co-culture growth medium described above. Some Lrp4 beads were inadvertently removed during washing, which may explain the absence of beads at some synapsin clusters. Uptake of the styryl dye FM4-64FX, a tracer for recycling synaptic vesicles, was assessed by incubating cells for 2 min in a depolarizing buffer (90 mM KCl, 64 mM NaCl, 2 mM $\mathrm{CaCl_2}$, 2 mM $\mathrm{MgCl_2}$, 10 mM glucose and 20 mM HEPES, pH 7.2) containing 10 μM FM4-64FX (Invitrogen). After washing in a non-depolarizing buffer, dye release was monitored by depolarizing cells further for 2 min, mainly as described previously[29,30]. Images were acquired on a Zeiss 510 confocal microscope and analysed using Volocity 3D imaging software (Perkin Elmer). We defined synaptic puncta as synapsin clusters that were $\geq 3\ \mu\mathrm{m}^2$ in size for co-cultures of motor neurons and HEK 293 or NIH 3T3 cells and $\geq 1.5\ \mu\mathrm{m}^2$ in size for motor neurons treated with microspheres. FM4-64FX clusters that were $\geq 1.5\ \mu\mathrm{m}^2$ in size were designated as puncta. We determined the number of puncta in a field of $1.44 \times 10^4\ \mu\mathrm{m}^2$.

Staining with AP–Lrp4. AP–Lrp4 fusion proteins were generated as described previously[14], and their concentrations were determined by measuring alkaline phosphatase activity. Explants were incubated for 90 min at room temperature (23–28 °C) with culture medium containing alkaline phosphatase fusion proteins (10 nM) in binding buffer (150 mM NaCl, 2 mM $\mathrm{CaCl_2}$, 1 mM $\mathrm{MgCl_2}$, 0.2% BSA and 20 mM HEPES, pH 7.2, 0.1% $\mathrm{NaN_3}$). After washing 5 times in binding buffer, the explants were fixed for 10 min in 3.7% formaldehyde, washed three times in HBS (HEPES-buffered saline) (150 mM NaCl and 20 mM HEPES, pH 7.2) and incubated for 30 min at 65 °C to inactivate endogenous alkaline phosphatase activity. After 3 washes in reaction buffer (100 mM NaCl, 50 mM $\mathrm{MgCl_2}$ and 100 mM Tris, pH 9.5), alkaline phosphatase activity was revealed by overnight incubation in reaction buffer with NBT/BCIP (Roche) at room temperature (23–28 °C). Images were acquired with a charge-coupled device (CCD) camera (Princeton Instruments) and were analysed with MetaMorph or Image J. To quantitate binding along the proximal–distal axis, we measured staining along short axon segments at varying distances from the soma, and we subtracted the values for binding of alkaline phosphatase alone from the values for AP–ecto-Lrp4.

Mice. Mice that are agrin-null, mutant for *Lrp4* (*mitt*), or carry an *actin*::*Musk* transgene (*Musk-L*), which increases *Musk* expression by threefold, have been described previously[3,9,18]. Similar results were found in mice that overexpress *Musk* by 20-fold (*Musk-H*)[18]. Diaphragm muscles from E18.5 mice were dissected and stained with antibodies to neurofilament and synaptophysin, and with α-BGT, as described previously[18]. We examined at least 70 AChR clusters from 3 or more mice of each genotype.

28. Hata, K., Polo-Parada, L. & Landmesser, L. T. Selective targeting of different neural cell adhesion molecule isoforms during motoneuron myotube synapse formation in culture and the switch from an immature to mature form of synaptic vesicle cycling. *J. Neurosci.* **27,** 14481–14493 (2007).
29. Umemori, H., Linhoff, M. W., Ornitz, D. M. & Sanes, J. R. FGF22 and its close relatives are presynaptic organizing molecules in the mammalian brain. *Cell* **118,** 257–270 (2004).
30. Umemori, H. & Sanes, J. R. Signal regulatory proteins (SIRPS) are secreted presynaptic organizing molecules. *J. Biol. Chem.* **283,** 34053–34061 (2008).

doi:10.1038/nature11334

Genome-wide association study indicates two novel resistance loci for severe malaria

Christian Timmann[1,2], Thorsten Thye[1,2], Maren Vens[2], Jennifer Evans[1,3], Jürgen May[4], Christa Ehmen[1], Jürgen Sievertsen[1], Birgit Muntau[1], Gerd Ruge[1], Wibke Loag[4], Daniel Ansong[5], Sampson Antwi[5], Emanuel Asafo-Adjei[5], Samuel Blay Nguah[5], Kingsley Osei Kwakye[5], Alex Osei Yaw Akoto[5], Justice Sylverken[5], Michael Brendel[1,2], Kathrin Schuldt[1], Christina Loley[2], Andre Franke[6], Christian G. Meyer[1], Tsiri Agbenyega[5], Andreas Ziegler[2] & Rolf D. Horstmann[1]

Malaria causes approximately one million fatalities per year, mostly among African children[1]. Although highlighted by the strong protective effect of the sickle-cell trait[2,3], the full impact of human genetics on resistance to the disease remains largely unexplored[4]. Genome-wide association (GWA) studies are designed to unravel relevant genetic variants comprehensively; however, in malaria, as in other infectious diseases, these studies have been only partly successful[5]. Here we identify two previously unknown loci associated with severe falciparum malaria in patients and controls from Ghana, West Africa. We applied the GWA approach to the diverse clinical syndromes of severe falciparum malaria, thereby targeting human genetic variants influencing any step in the complex pathogenesis of the disease. One of the loci was identified on chromosome 1q32 within the *ATP2B4* gene, which encodes the main calcium pump of erythrocytes[6], the host cells of the pathogenic stage of malaria parasites. The second was indicated by an intergenic single nucleotide polymorphism on chromosome 16q22.2, possibly linked to a neighbouring gene encoding the tight-junction protein MARVELD3. The protein is expressed on endothelial cells[7] and might therefore have a role in microvascular damage caused by endothelial adherence of parasitized erythrocytes. We also confirmed previous reports on protective effects of the sickle-cell trait and blood group O[5,8,9]. Our findings underline the potential of the GWA approach to provide candidates for the development of control measures against infectious diseases in humans.

Malaria is caused by the protozoan parasites *Plasmodium falciparum* and less virulent plasmodia. The severe form of the disease, severe falciparum malaria (SM), may comprise distinct or overlapping clinical syndromes including severe anaemia, cerebral malaria presenting as coma and/or convulsions, acidosis, respiratory distress, prostration, and several less frequent complications[10].

In a GWA study we included 1,325 SM cases of severe anaemia or cerebral malaria as well as 828 unaffected controls. To corroborate the phenotype of SM, only cases with concomitant acidosis and/or respiratory distress were enrolled[11]. The Affymetrix Genome-Wide Human SNP Array 6.0 was used for genotyping. After stringent quality control (Supplementary Methods), genome-wide imputation was performed with the MACH software (version 1.0.16) using genotype data for 174 individuals of African descent who were included in the 2010-08 release of the 1,000 Genomes Project (1000G; http://www.1000genomes.org; for details see Supplementary Methods), providing a total of 5,010,634 single nucleotide polymorphisms (SNPs) for further analysis. Assuming an additive mode of inheritance (MOI) and adjusting for age and gender and for population stratification in a logistic regression model with the first three components of a principal-components analysis (PCA; Supplementary Methods and

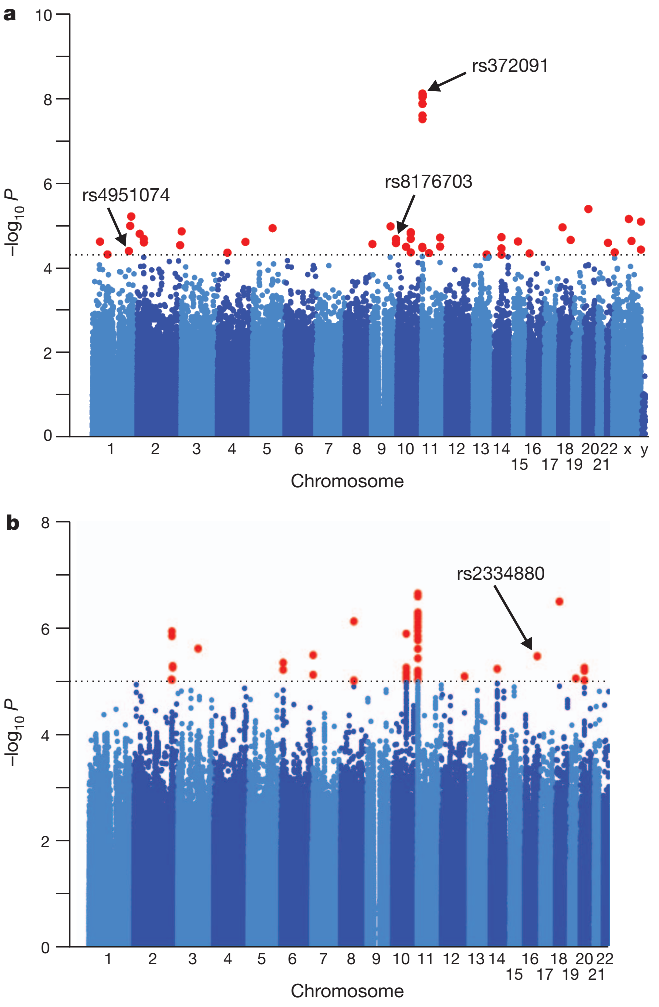


Figure 1 | Manhattan plots for GWA with severe malaria. a, b, GWA screen with 804,895 genotyped SNPs (**a**) and 4,205,739 imputed SNPs (**b**) in 2,153 individuals. Values of $-\log_{10} P$ are plotted against chromosomal positions. Red dots indicate SNPs with $P < 5 \times 10^{-5}$ (**a**) and $P < 10^{-5}$ (**b**); dotted lines indicate these thresholds. All other SNPs are given in blue. Signals reaching genome-wide significance in the joint analysis after replication are indicated by arrows and SNP numbers. **a**, rs4951074 at chromosome 1q32.1, *ATP2B4* locus; rs8176719 at chromosome 9q34.2, *ABO* locus; rs372091 at chromosome 11p15.5, *HBB* locus. **b**, rs2334880 at chromosome 16q22.2, *MARVELD3* locus.

[1]Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, 20359 Hamburg, Germany. [2]Institute of Medical Biometry and Statistics, University at Lübeck, 23562 Lübeck, Germany. [3]Kumasi Centre for Collaborative Research in Tropical Medicine, Kumasi, Ghana. [4]Infectious Disease Epidemiology Group, Bernhard Nocht Institute for Tropical Medicine, 20359 Hamburg, Germany. [5]School of Medical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [6]Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, 24105 Kiel, Germany.

Table 1 | Association signals of four loci for severe falciparum malaria

				Ghanaian GWA group (n = 2,153)		Ghanaian replication group (n = 3,542)		Combined Ghanaian groups	Gambian GWA group (n = 2,213)		Meta-analysis, Ghana, Gambia
Gene Locus MOI	Chr. bp position (hg19)	SNP, (major allele/ minor allele)	MAF (Caucasian ref. panels)	MAF, cases (ctrl)	OR (95% CI) *P*	MAF, cases (ctrl)	OR (95% CI) *P*	OR (95% CI) *P*	MAF, cases (ctrl)	OR (95% CI) *P*	OR *P*
ATP2B4 1q32.1 Rec	203654024	rs10900585 (T/G), identified by imputation	0.15 (CEU)	0.38 (0.43)	0.60 (0.47–0.76) 2.0×10^{-5}	0.38 (0.44)	0.69 (0.57–0.84) 1.9×10^{-4}	0.65 (0.56–0.75) 6.1×10^{-9}	0.33* (0.37)*	0.58* (0.40–0.85) 5.2×10^{-3}*	0.61 1.9×10^{-10}
	203656230	rs2365860 (A/C), GWA marker	0.088 (CEU)	0.33 (0.39)	0.58 (0.45–0.74) 2.5×10^{-5}	0.33 (0.39)	0.68 (0.55–0.84) 3.9×10^{-4}	0.63 (0.55–0.74) 1.5×10^{-8}	0.27 (0.30)	0.54 (0.36–0.81) 2.7 x 10^{-3}	0.61 1.9×10^{-10}
	203656974	rs10900589 (T/A), GWA marker	0.088 (CEU)	0.33 (0.39)	0.57 (0.44–0.74) 2.4×10^{-5}	0.33 (0.39)	0.69 (0.55–0.85) 5.1×10^{-4}	0.63 (0.54–0.74) 2.1×10^{-8}	0.27 (0.29)	0.54 (0.36–0.81) 2.8×10^{-3}	0.62 2.8×10^{-10}
	203657749	rs2365858 (C/G), GWA marker	0.094 (CEU)	0.33 (0.39)	0.56 (0.43–0.73) 1.1×10^{-5}	0.33 (0.38)	0.68 (0.55–0.85) 5.9×10^{-4}	0.63 (0.54–0.74) 5.1×10^{-8}	0.27 (0.29)	0.56 (0.37–0.83) 4.4×10^{-3}	0.62 9.5×10^{-10}
	203660781	rs4951074 (G/A), GWA marker	0.088 (CEU)	0.32 (0.38)	0.55 (0.42–0.71) 6.5×10^{-6}	0.32 (0.37)	0.66 (0.53–0.83) 2.6×10^{-4}	0.62 (0.53–0.74) 3.4×10^{-8}	0.29 (0.30)	0.61 (0.42–0.89) 1.1×10^{-2}	0.62 1.3×10^{-9}
ABO 9q34.2 Dom	136132908	rs8176719 (delG/G), causal variant, Ala87fs	0.41 (PGA)	0.36 (0.29)	1.62 (1.35–1.96) 1.2×10^{-7}	0.37 (0.28)	1.70 (1.48–1.96) 2.9×10^{-13}	1.67 (1.50–1.86) 1.1×10^{-20}	n.a. (0.16)†	1.26 † (1.11–1.44) 5×10^{-4}†	1.48 4.3×10^{-21}
	136135863	rs8176703 (C/A), GWA marker	0.00 (CEU)	0.093 (0.057)	1.84 (1.42–2.39) 5.1×10^{-6}	0.091 (0.056)	1.72 (1.41–2.10) 1.3×10^{-7}	1.73 (1.48–2.02) 4.0×10^{-12}	n.c.‡	n.c.‡	n.c.
HBB 11p15.5 Het	5248232	rs334 (A/T), causal variant (Glu6Val/HbS)	0.00 (CEU)	0.0053 (0.072)	0.066 (0.038–0.12) 2.5×10^{-21}	0.010 (0.059)	0.15 (0.10–0.23) 1.6×10^{-18}	0.11 (0.079–0.15) 1.4×10^{-38}	n.a.	n.a. 1.3×10^{-28}†	n.c.
	5518156	rs372091 (C/T), GWA marker	0.00 (CEU)	0.029 (0.068)	0.38 (0.28–0.52) 1.4×10^{-9}	0.028 (0.058)	0.45 (0.34–0.59) 3.6 x 10^{-8}	0.44 (0.36–0.54) 1.1×10^{-14}	0.012 (0.018)	0.63 (0.38–1.07) 0.085	0.46 5.6×10^{-14}
MARVELD3 16q22.2 Add	71653637	rs2334880 (T/C), identified by imputation	0.29 (CEU)	0.47 (0.40)	1.31 (1.16–1.49) 2.3×10^{-5}	0.45 (0.41)	1.20 (1.08–1.32) 4.3×10^{-4}	1.24 (1.15–1.34) 3.9×10^{-8}	0.40* (0.38)*	0.96* (0.81–1.13) 0.60*	1.19 1.9×10^{-6}

Results for all SNPs are based on physical genotyping unless otherwise indicated. For *ATP2B4* and *MARVELD3* variants the most plausible MOI was derived from both a Max test and logistic regression (Supplementary Table 4). MOIs for *ABO* and *HBB* were taken from refs 8, 22 and 23. hg19, position according to the Genome Reference Consortium Human Build 37.3 (http://www.ncbi.nlm.nih.gov/genome/assembly/); Rec, recessive; Dom, dominant; Het, heterozygous advantage; Add, additive; CEU, Caucasian HapMap panel; PGA, Programs for Genomic Applications-European Panel; MAF, minor allele frequency; del, deletion; fs, frame shift; n.a., data not available; n.c., not calculated. *Results calculated after imputation of the 1000G data into the Gambian data set. †Data taken from ref. 5. ‡Imputation failed (MACH RSQ <0.3; see Supplementary Discussion).

Supplementary Fig. 1), we identified 102 SNPs located in 41 distinct genomic regions (Fig. 1) using thresholds of $P < 5 \times 10^{-5}$ in the GWA data and $P < 10^{-5}$ in the imputation data. The more stringent threshold of $P < 10^{-5}$ was applied to the imputed data because the African genotypes of 1000G were derived from different ethnicities, and this may decrease imputation accuracy. The genomic inflation factor λ of the P values of the GWA study was estimated to be 1.045 for physically typed SNPs and 1.043 for imputed SNPs (Supplementary Fig. 2). The genotypes of the lead SNPs of each of the 41 genomic regions were validated by repeat genotyping, and the initial association signals were confirmed in 40 regions (Supplementary Table 1).

Subsequently, replication experiments were performed in an additional 1,320 SM cases and 2,222 controls from the same population (Supplementary Table 1). Statistical evaluation was based on a logistic regression model including the variables of age, gender and self-reported ethnicities.

In the joint analysis, four loci showed $P < 5 \times 10^{-8}$, considered to denote genome-wide significance[12,13] (Supplementary Table 2). The associations were supported when the various ethnic groups included in the study were tested separately (Supplementary Table 3).

Two of the four loci are novel. The first locus was identified by the genome-wide array. It is indicated by SNP rs4951074 on chromosome 1q32.1 (Fig. 1a). Several SNPs in this region reached genome-wide significance (Fig. 2a and Table 1), and the most likely MOI is recessive (Supplementary Table 4). These SNPs are all located within *ATP2B4* (ATPase, Ca^{2+}-transporting, plasma membrane, 4; MIM ID *108732 (Online Mendelian Inheritance in Man; http://www.ncbi.nlm.nih.gov/)), the gene encoding the plasma membrane Ca^{2+}-ATPase type 4 (PMCA4).

Two approaches were used to search for the causal variant. First, from SNPs imputed on chromosome 1 we selected those that showed a linkage disequilibrium to the lead SNP rs4951074 of $r^2 > 0.5$ or a linkage disequilibrium of $r^2 > 0.05$ together with $P < 10^{-3}$ for association based on a recessive MOI. Second, all *ATP2B4* exons and a 1,000-base-pair upstream region were screened for additional coding or promoter variants by high-resolution melting and resequencing. Together this resulted in 73 SNPs, which were genotyped for fine mapping. SNP rs10900585 showed the strongest association in the joint Ghanaian study groups (odds ratio (OR) = 0.65, 95% confidence interval (CI) = 0.56–0.75, $P = 6.1 \times 10^{-9}$; Fig. 2a and Table 1). It is located in the second intron of *ATP2B4* and has no apparent function. The haplotype block covered roughly 500 kilobases (kb), as derived from linkage disequilibrium values of $r^2 > 0.05$ between rs10900585 and SNPs genotyped or imputed on chromosome 1 (Supplementary Fig. 3).

An independent replication group was made available through data provided by the MalariaGEN Network of a GWA study on SM among Gambian children[5]. All four signal SNPs of the Ghanaian WGA group that were also genotyped in the Gambian study, as well as SNP rs10900585 imputed from the 1000G data into the Gambian genotypes, had the same risk alleles and similar ORs to those in our study and $P < 0.05$ (Table 1).

The localization of several associated SNPs inside *ATP2B4* suggests that variants of *ATP2B4* itself and its product PMCA4 do indeed contribute to susceptibility and resistance to SM. PMCA4 is a ubiquitous Ca^{2+} pump that occurs in tissue-specific splice variants[14]; genetic

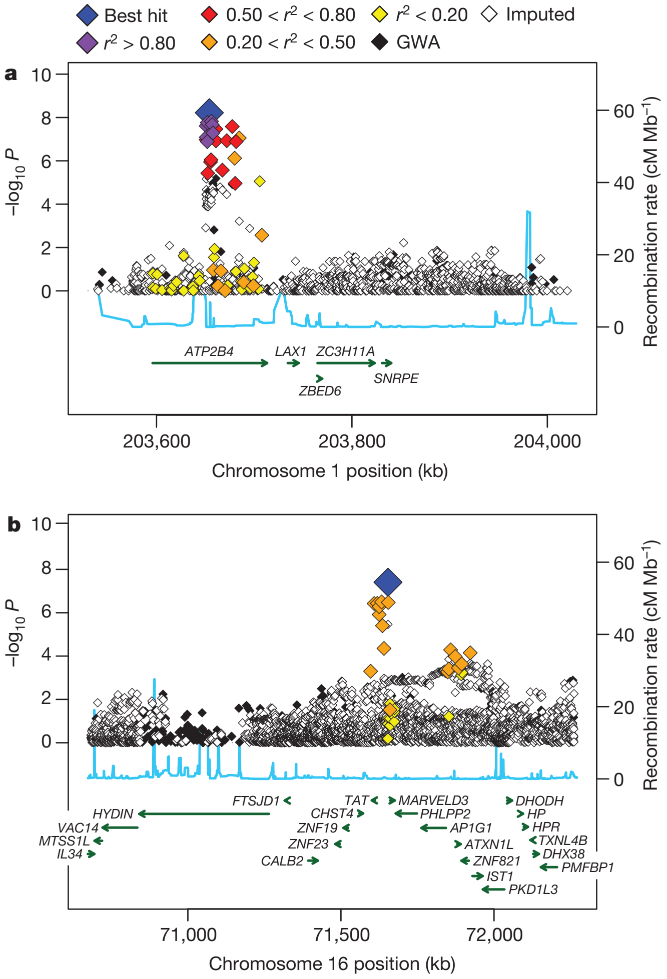


Figure 2 | Regional association plots for new loci at 1q32.1 and 16q22.2. **a**, **b**, *ATP2B4* locus (**a**) and *MARVELD3* locus (**b**) as defined by the positions of SNPs showing a linkage disequilibrium of $r^2 = 0.05$ with the lead SNP rs10900585 (**a**; $P = 6.1 \times 10^{-9}$) and SNP rs2334880 (**b**; $P = 3.9 \times 10^{-8}$), respectively. Disease associations as indicated by $-\log_{10}$ *P*-values are plotted against chromosomal positions. Black and white diamonds represent individual SNPs of the GWA screen using genotyped and imputed data, respectively. Coloured diamonds indicate SNP data obtained by the analysis of the total of 2,645 SM cases and 3,050 controls. Additional SNPs selected for fine mapping are included (Supplementary Methods). Associations were assessed assuming recessive and additive modes of inheritance for the *ATP2B4* locus and the *MARVELD3* locus, respectively. Levels of linkage disequilibria (r^2) with the best-associated SNP (blue diamonds) are colour-coded. Blue lines indicate recombination fractions in accordance with the HapMap Caucasian panel sample. Horizontal arrows mark structural human genes as annotated by Human Genome Build 37.3/gh19 of the UCSC (Genome Bioinformatics Group, University of California, Santa Cruz; http://genome.ucsc.edu/cgi-bin/hgTracks).

variation may therefore influence either the expression or subtle or even gross structural properties of the molecule. Because PMCA4 is the major Ca^{2+} pump of erythrocytes[6], alteration of its structure or expression may disturb the homeostasis of intra-erythrocytic Ca^{2+} concentrations. Thus, it could affect the development and structure of the intra-erythrocytic stages of the parasite. To this end, it has been shown that a decrease in Ca^{2+} concentrations in the compartment separating the parasite from its host erythrocyte, the parasitophorous vacuole, may result in massive impairment of parasite reproduction and maturation[15]. PMCA4 forms of platelets and endothelial cells may be affected as well. Both cells are activated by intracellular Ca^{2+} (refs 16, 17). Activation of platelets has been reported to stimulate parasite killing[18], whereas activation of endothelial cells enhances vascular adherence of *P. falciparum*-infected erythrocytes[19], which is a key event in the pathogenesis of SM[2]. Taken together, several Ca^{2+}-dependent cellular functions[14] make PMCA4 a candidate for functional studies and, possibly, for future intervention. Deactivation of endothelial cells has been proposed as a supportive treatment for cerebral malaria[20].

The second novel locus, which is located on chromosome 16q22.2, was indicated by imputation and supported by physical genotyping of SNP rs2334880 (OR = 1.24, 95% CI = 1.15–1.34, $P = 3.9 \times 10^{-8}$ in the joint analysis; Figs 1b and 2b and Table 1). The haplotype block inferred from SNPs with a linkage disequilibrium of $r^2 > 0.05$ to rs2334880 spans 1,500 kb (Supplementary Fig. 3). Replication was attempted in the Gambian study group after imputation of the 1000G data. However, this association was not replicated in the Gambian samples (Table 1). This may be explained by the absence of the resistance allele from the Gambian population or, alternatively, by a difference between the haplotypes of the Gambian and Ghanaian populations, which, indeed, is indicated by the linkage disequilibrium patterns in the two study groups (Supplementary Fig. 4).

SNP rs2334880 maps to the intergenic region between *TAT* (tyrosine aminotransferase; MIM ID *613018) and *MARVELD3* (MARVEL domain-containing protein 3 gene; MIM ID *614094), which lie in a head-to-head configuration (Fig. 2b). rs2334880 is located 42.6 kb upstream of *TAT* and 6.4 kb upstream of *MARVELD3*. Whereas *TAT* encodes a housekeeping enzyme, the gene product of *MARVELD3* is part of tight-junction structures of epithelial and vascular endothelial cells[7]. Its function in the endothelium seems to be of interest because endothelial adherence of infected erythrocytes is important in the pathology of SM[2,3]. Structural variants of the MARVELD3 protein or alterations in its expression could influence the barrier function and inflammatory reactivity of the endothelium[21] and, thereby, the course of the disease. Fine mapping performed as described above for the *ATP2B4* locus included 35 SNPs, none of which showed a stronger disease association than rs2334880 (Fig. 2b). It is conceivable, however, that the causal variant remains to be determined and that it might act on a more distant gene.

Two additional association signals were obtained in our study, confirming earlier findings[8] (Fig. 1a, Table 1, Supplementary Fig. 3 and Supplementary Discussion). The strongest signal was caused by the sickle-cell allele haemoglobin S (HbS) at the β-globin gene *HBB*, which is in agreement with the result of the Gambian GWA study[5]. A further signal came from SNPs located in the *ABO* gene and supports previous reports indicating that blood group O has some protective effect against severe malaria[5,22,23].

With reference to a recent controversial discussion[24,25], our findings underline the value of GWA studies as a systematic approach for identifying molecular determinants relevant to protection from infectious diseases. Both new association signals reported here lend themselves to a straightforward elaboration by cellular biology and, possibly, to medical application.

METHODS SUMMARY

Study group. SM was diagnosed in accordance with the definition of the World Health Organization[10]. Specifically, cerebral malaria was defined by a Blantyre coma score[26] of less than 3 in the presence of *P. falciparum* parasitaemia and severe anaemia by a haemoglobin concentration of less than 5 g dl^{-1} in the presence of *P. falciparum* parasitaemia. Further definitions of SM are provided in Supplementary Methods.

Genotyping. Genotyping was performed with the Affymetrix Genome-Wide Human SNP Array 6.0. For replication and fine mapping, SNPlex technology, DNA resequencing (Applied Bioscience) and hybridization assays were used. SNP detection was performed by resequencing in 16 individuals as well as by high-resolution melting analyses in 350 individuals and resequencing where appropriate (Supplementary Table 5).

Statistics. Associations were tested by using a logistic regression model (software PLINK v. 1.07; ref. 27). Excluded were individuals with whole-genome amplified DNA, with autosomal heterozygosity outside the range 26–31%, with genotype-call rates less than 96%, or with identity by descent at least 0.125 to other study participants as well as SNPs with minor allele frequencies (MAFs) <1%

or genotyping rates less than 96%. After quality control, 804,895 informative SNPs were available for the analysis in the Ghanaian study group. Genome-wide imputation of the African 1000G data yielded 4,205,739 SNPs with MAFs of more than 10% and an imputation quality score of R-squared (RSQ) values greater than 0.8 (Minimac software). Imputations on chromosomes 1, 9, 11 and 16 of the Ghanaian WGA group for fine mapping and of the Gambian WGA study for replication attempts were performed with MAFs of more than 5% and RSQ values greater than 0.3. The joint analysis of the Ghanaian GWA study group and the Ghanaian replication group was performed by merging both data sets using genotypes determined by SNPlex or hybridization assays. A fixed-effects meta-analysis of the Ghanaian and Gambian studies was calculated (PLINK).

Received 24 November 2011; accepted 18 June 2012.

Published online 15 August 2012.

1. Murray, C. J. L. *et al.* Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet* **379,** 413–431 (2012).
2. Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. The pathogenic basis of malaria. *Nature* **415,** 673–679 (2002).
3. Idro, R., Marsh, K., John, C. C. & Newton, C. R. Cerebral malaria: mechanisms of brain injury and strategies for improved neurocognitive outcome. *Pediatr. Res.* **68,** 267–274 (2010).
4. Mackinnon, M. J., Mwangi, T. W., Snow, R. W., Marsh, K. & Williams, T. N. Heritability of malaria in Africa. *PLoS Med.* **12,** e340 (2005).
5. Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genet.* **41,** 657–665 (2009).
6. Stauffer, T. P., Guerini, D. & Carafoli, E. Tissue distribution of the four gene products of the plasma membrane Ca^{2+} pump. A study using specific antibodies. *J. Biol. Chem.* **270,** 12184–12190 (1995).
7. Steed, E., Rodrigues, N. T., Balda, M. S. & Matter, K. Identification of MarvelD3 as a tight junction-associated transmembrane protein of the occludin family. *BMC Cell Biol.* **10,** 95 (2009).
8. Kwiatkowski, D. P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77,** 171–192 (2005).
9. May, J. *et al.* Hemoglobin variants and disease manifestations in severe falciparum malaria. *J. Am. Med. Assoc.* **297,** 2220–2226 (2007).
10. World Health Organization. Communicable diseases cluster: severe falciparum malaria. *Trans. R. Soc. Trop. Med. Hyg.* **94** (suppl. 1), S1–S90 (2000).
11. Marsh, K. *et al.* Indicators of life-threatening malaria in African children. *N. Engl. J. Med.* **332,** 1399–1404 (1995).
12. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32,** 381–385 (2008).
13. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
14. Brini, M. & Carafoli, E. Calcium pumps in health and disease. *Physiol. Rev.* **89,** 1341–1378 (2009).
15. Gazarini, M. L., Thomas, A. P., Pozzan, T. & Garcia, C. R. Calcium signaling in a low calcium environment: how the intracellular malaria parasite solves the problem. *J. Cell Biol.* **161,** 103–110 (2003).
16. Szewczyk, M. M. *et al.* Ca^{2+}-pumps and Na^{2+}–Ca^{2+}-exchangers in coronary artery endothelium versus smooth muscle. *J. Cell. Mol. Med.* **11,** 129–138 (2007).
17. Varga-Szabo, D., Braun, A. & Nieswandt, B. Calcium signaling in platelets. *J. Thromb. Haemost.* **7,** 1057–1066 (2009).
18. McMorran, B. J. *et al.* Platelets kill intraerythrocytic malarial parasites and mediate survival to infection. *Science* **323,** 797–800 (2009).
19. Bridges, D. J. *et al.* Rapid activation of endothelial cells enables *Plasmodium falciparum* adhesion to platelet-decorated von Willebrand factor strings. *Blood* **115,** 1472–1474 (2010).
20. Wassmer, S. C., Cianciolo, G. J., Combes, V. & Grau, G. E. Inhibition of endothelial activation: a new way to treat cerebral malaria? *PLoS Med.* **2,** e245 (2005).
21. Cinel, I. & Dellinger, R. P. Advances in pathogenesis and management of sepsis. *Curr. Opin. Infect. Dis.* **20,** 345–352 (2007).
22. Loscertales, M. *et al.* ABO blood group phenotypes and *Plasmodium falciparum* malaria: unlocking a pivotal mechanism. *Adv. Parasitol.* **65,** 2–41 (2007).
23. Fry, A. E. *et al.* Common variation in the ABO glycosyltransferase is associated with susceptibility to severe *Plasmodium falciparum* malaria. *Hum. Mol. Genet.* **17,** 567–576 (2008).
24. Editorial. Are genome-wide association studies of infection any value? *Lancet Infect. Dis.* **10,** 577 (2010).
25. de Bakker, P. I. & Telenti, A. Infectious diseases not immune to genome-wide association. *Nature Genet.* **42,** 731–732 (2010).
26. Molyneux, M. E., Taylor, T. E., Wirima, J. J. & Borgstein, A. Clinical features and prognostic indicators in paediatric cerebral malaria: a study of 131 comatose Malawian children. *Q. J. Med.* **71,** 441–459 (1989).
27. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the participating children, their parents and guardians; L. N. Badu, S. Opoku, M. Attan-Ayibo and D. Sambian for technical assistance. Ethical approval was obtained from the Committee for Research, Publications and Ethics of the School of Medical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. Informed consent was obtained from the parents or guardians of the children and documented by signature, or thumbprint in cases of illiteracy. The study complied with the Ethical Principles for Medical Research Involving Human Subjects as laid out in the Declaration of Helsinki. The study was supported by the National Genome Research Network (NGFN1, NGFN2) of the German Ministry for Education and Research (BMBF). In addition, this study makes use of data generated by MalariaGEN[5]. A full list of the investigators who contributed to the generation of the data is available from http://www.malariagen.net. The MalariaGEN Project is supported by the Wellcome Trust (WT077383/Z/05/Z) and by the Foundation for the National Institutes of Health (566) as part of the Bill & Melinda Gates' Grand Challenges in Global Health Initiative.

Author Contributions R.D.H., C.T. and A.Z. designed the study. C.T., R.D.H., C.G.M., A.Z. and T.T. drafted the manuscript. D.A., S.A., E.A.A., S.B.N., K.O.K., A.O.Y.A., J.Sy., W.L., J.E. and J.M. recruited cases and controls and acquired materials. C.E., J.Si., B.M. and G.R. performed resequencing and SNP genotyping. T.T., C.T., M.B., M.V., K.S., C.L., A.Z., J.M. and R.D.H. calculated the statistical analysis and/or interpreted data. R.D.H. and A.Z. obtained funding. J.E., T.A., J.M. and A.F. provided administrative, technical or material support. R.D.H., A.Z. and T.A. supervised the study.

Author Information Genetic variants identified by Sanger-based resequencing at the *ATP2B4* and *MARVELD3* loci are deposited in dbSNP (see Supplementary Table 5). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.T. (timmann@bnitm.de).

LETTER

doi:10.1038/nature11352

A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality

Bethany A. Buckley[1]*, Kirk B. Burkhart[1]*, Sam Guoping Gu[2], George Spracklin[1], Aaron Kershner[3], Heidi Fritz[1], Judith Kimble[1,3], Andrew Fire[2] & Scott Kennedy[1]

Epigenetic information is frequently erased near the start of each new generation[1]. In some cases, however, epigenetic information can be transmitted from parent to progeny (multigenerational epigenetic inheritance)[2]. A particularly notable example of this type of epigenetic inheritance is double-stranded RNA-mediated gene silencing in *Caenorhabditis elegans*. This RNA-mediated interference (RNAi) can be inherited for more than five generations[3–8]. To understand this process, here we conduct a genetic screen for nematodes defective in transmitting RNAi silencing signals to future generations. This screen identified the *heritable RNAi defective 1* (*hrde-1*) gene. *hrde-1* encodes an Argonaute protein that associates with small interfering RNAs in the germ cells of progeny of animals exposed to double-stranded RNA. In the nuclei of these germ cells, HRDE-1 engages the nuclear RNAi defective pathway to direct the trimethylation of histone H3 at Lys 9 (H3K9me3) at RNAi-targeted genomic loci and promote RNAi inheritance. Under normal growth conditions, HRDE-1 associates with endogenously expressed short interfering RNAs, which direct nuclear gene silencing in germ cells. In *hrde-1*- or nuclear RNAi-deficient animals, germline silencing is lost over generational time. Concurrently, these animals exhibit steadily worsening defects in gamete formation and function that ultimately lead to sterility. These results establish that the Argonaute protein HRDE-1 directs gene-silencing events in germ-cell nuclei that drive multigenerational RNAi inheritance and promote immortality of the germ-cell lineage. We propose that *C. elegans* use the RNAi inheritance machinery to transmit epigenetic information, accrued by past generations, into future generations to regulate important biological processes.

We conducted a genetic screen to identify factors required for multigenerational RNAi inheritance. We mutagenized animals carrying a germline green fluorescent protein (*gfp*) reporter gene, and screened for mutant animals that retained the ability to silence *gfp* when exposed directly to *gfp* double-stranded RNA (dsRNA), but failed to silence *gfp* in subsequent generations. Among fourteen mutant alleles fulfilling these criteria, four alleles defined a gene we term here *heritable RNAi defective 1* (*hrde-1*) (Supplementary Fig. 2). *hrde-1* mutant animals silenced GFP expression when exposed to *gfp* dsRNA, but failed to transmit this silencing to subsequent generations (Fig. 1a and Supplementary Fig. 3). Similarly, *hrde-1* mutants silenced the germline-expressed *oma-1* gene when treated directly with *oma-1* dsRNA, but were defective for silencing inheritance (Supplementary Fig. 4). We conclude that *hrde-1* promotes multigenerational RNAi silencing in the germ line.

We mapped *hrde-1* to a genomic region containing the *c16c10.3* (also known as *wago-9*) gene, which is predicted to encode an Argonaute (AGO) protein not known to contribute to gene silencing[9,10] (see note added in proof). *c16c10.3* encodes a predicted bipartite nuclear localization signal (NLS) and PAZ and PIWI domains (Supplementary Fig. 5). We found that *hrde-1* is *c16c10.3* (Supplementary Fig. 5), and a member of the worm-specific clade of AGO (WAGO) family[9]. HRDE-1 seemed to be relatively unique among the WAGO proteins in its contribution to germline RNAi inheritance (Supplementary Fig. 6). We constructed a fusion gene between *gfp* and a full-length genomic copy of *hrde-1* (*gfp::hrde-1*). *gfp::hrde-1* rescued RNAi inheritance in *hrde-1*$^{-/-}$ animals, indicating that GFP–HRDE-1 is functional (Supplementary Fig. 5). GFP–HRDE-1 was expressed in the nuclei of male and female germ cells (Fig. 1b and Supplementary Fig. 7). These data indicate that *hrde-1* encodes a germline AGO that localizes to the nucleus.

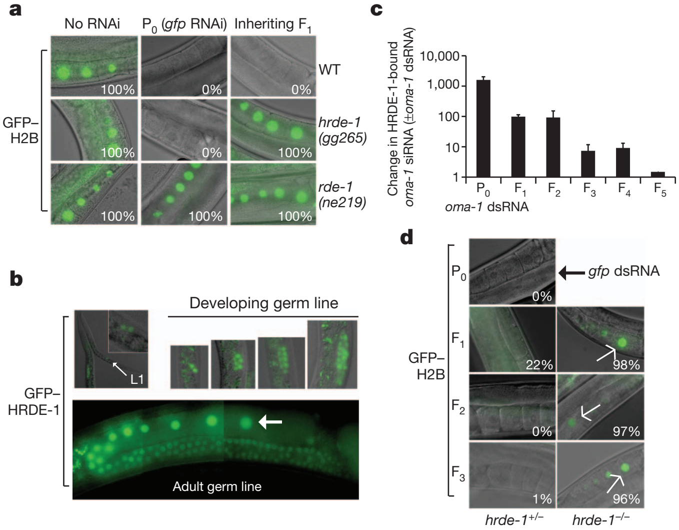


Figure 1 | *hrde-1* encodes a nuclear AGO protein that acts in inheriting generations to promote multigenerational germline RNAi inheritance. **a**, *pie-1::gfp::h2b*-expressing animals were exposed to *gfp* dsRNA. F_1 progeny were grown in the absence of dsRNA, and GFP expression in oocytes was visualized by fluorescence microscopy. The percentage of fluorescent animals is indicated. *rde-1* is required for RNAi[17]. P_0, parental generation; WT, wild-type ($n > 100$). **b**, GFP–HRDE-1 was visualized by fluorescent microscopy. Small arrow, L1 animal; inset, magnifications showing GFP–HRDE-1 in primordial germ cells. Large arrow, oocyte nucleus. **c**, 3×Flag–HRDE-1 was immunoprecipitated with an anti-Flag antibody and co-precipitating *hrde-1* RNA was isolated, and *oma-1* siRNAs were quantified with an *oma-1* TaqMan probe set. Data are expressed as fold change (± *oma-1* RNAi), '1' denotes no change, and the mean and s.d. are shown ($n = 3$). **d**, *pie-1::gfp::h2b* fluorescence was scored in *hrde-1*$^{+/-}$ and *hrde-1*$^{+/+}$ (scored as one group) or in *hrde-1*$^{-/-}$ inheriting progeny. The *hrde-1*$^{-/-}$ chromosome was marked with *unc-93* (*unc-93* is ~1.3 cM from *hrde-1*), and *hrde-1* genotypes were inferred by Unc phenotypes. The percentage of fluorescent animals is indicated. Original magnification for all images, ×630.

[1] Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [2] Departments of Pathology and Genetics, Stanford University, Stanford, California 94305, USA. [3] Howard Hughes Medical Institute, and Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.

*These authors contributed equally to this work.

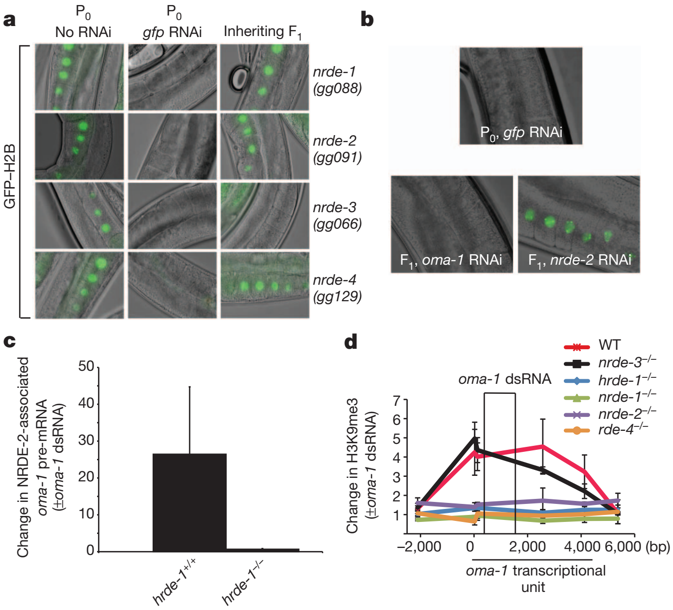


Figure 2 | HRDE-1 engages the NRDE nuclear RNAi pathway to direct multigenerational RNAi inheritance. a, *pie-1::gfp::h2b* fluorescence is shown. More than 98% of animals of each genotype exhibited phenotypes similar to that of the image shown. **b**, P_0 animals expressing *pie-1::gfp::h2b* were exposed to *gfp* dsRNA. F_1 progeny were exposed to *oma-1* or *nrde-2* dsRNA ($n = 3$). **c**, Flag–NRDE-2 was precipitated with an anti-Flag antibody and NRDE-2 co-precipitating *oma-1* pre-mRNA was quantified by rtPCR using exon/intron primer sets. Data are expressed as a ratio ± *oma-1* RNAi, and the mean and s.e.m. are shown ($n = 3$). **d**, The F_1 progeny of *oma-1* dsRNA-treated animals were subjected to H3K9me3 chromatin immunoprecipitation. Co-precipitating *oma-1* DNA was quantified by rtPCR. Data were normalized to co-precipitating *eft-3* DNA and expressed as a ratio ± *oma-1* RNAi (1 = no change). On the *x*-axis, 0 denotes the predicted start codon of *oma-1*. *rde-4* is required for RNAi[18]. Data are mean ± s.e.m. ($n = 3$–4). bp, base pair. Original magnification, ×630.

HRDE-1 could conceivably promote multigenerational RNAi inheritance by acting in animals directly exposed to dsRNA (the RNAi generation) or in the progeny of these animals (the inheriting generation). In *C. elegans*, dsRNA exposure induces the expression of short interfering RNAs (siRNAs) in inheriting generations (refs 7, 8 and Supplementary Fig. 8). HRDE-1 co-precipitated with siRNAs for several generations after exposure to RNAi, consistent with the idea that HRDE-1 acts in inheriting generations to promote RNAi inheritance (Fig. 1c). Note, the maintenance of HRDE-1-associating siRNA populations over generations is probably mediated by RNA-dependent RNA polymerases (RdRPs) (see Supplementary Discussion). The following genetic analyses confirmed that HRDE-1 acts in inheriting generations. Animals that were *hrde-1*$^{+/-}$ in the parental (P_0) RNAi generation, but were *hrde-1*$^{-/-}$ in the F_1 inheriting generation, failed to inherit RNAi silencing (Fig. 1d and Supplementary Table 1). Similarly, HRDE-1 activity was required in the F_2 generation for F_1-to-F_2 RNAi inheritance, and in the F_3 generation for F_2-to-F_3 RNAi inheritance (Fig. 1d). Conversely, animals that lacked HRDE-1 in the RNAi generation, but expressed HRDE-1 in the inheriting generation, were able to inherit RNAi silencing (Supplementary Table 1). Thus, HRDE-1 acts in inheriting progeny to facilitate the memory of RNAi silencing events that occurred in previous generations. Altogether, these data establish that *C. elegans* possess machinery dedicated to propagating epigenetic information across generational boundaries.

The nuclear RNAi defective factors 1–4 (NRDE-1–4) comprise a sub-branch of the *C. elegans* RNAi silencing machinery that is required for dsRNA-based silencing of nuclear-localized RNAs[11–13]. According to our current model, siRNAs bound to the somatically expressed AGO NRDE-3 recognize and bind nascent RNA transcripts and recruit NRDE-1, -2 and -4 (termed downstream NRDE factors) to genomic sites of RNAi in somatic cells. Together, the NRDE factors direct nuclear gene-silencing events, which include the deposition of the repressive chromatin mark H3K9me3, and the inhibition of RNA polymerase II elongation[11–13]. The NRDE factors contribute to heritable gene silencing events that are manifest in somatic cells[7]. Five lines of evidence indicate that HRDE-1 engages the downstream NRDE factors to direct nuclear RNAi, and, consequently, RNAi inheritance in germ cells. First, the downstream NRDE factors were required for *gfp* and *pos-1* germline RNAi inheritance (Fig. 2a and Supplementary Fig. 9). Second, like HRDE-1, the downstream NRDE factor NRDE-2 acted in inheriting generations to promote the memory of RNAi in germ cells (Fig. 2b). Third, HRDE-1 was required for RNAi-mediated recruitment of NRDE-2 to a germline pre-messenger RNA, indicating that HRDE-1 acts as a specificity factor in germ cells to recruit a downstream NRDE factor to genomic sites of RNAi (Fig. 2c). Fourth, the ability of dsRNA to induce H3K9me3 was lost in mutant strains that eliminate *hrde-1* or the downstream NRDE factors (Fig. 2d, Supplementary Fig. 10 and Supplementary Discussion). Fifth, consistent with the idea that *hrde-1* and the downstream NRDE factors act together in the germ line, *hrde-1*$^{-/-}$ animals share a germline mortality phenotype with *nrde-1/2/4*$^{-/-}$ animals (see later). These data indicate that NRDE-1, -2 and -4 are required for multigenerational RNAi inheritance, and support a model in which HRDE-1 and NRDE-1, -2 and -4 comprise a germline RNAi pathway that drives RNAi inheritance by inducing gene silencing in the nuclei of inheriting progeny. Henceforth, we refer to HRDE-1 and NRDE-1, -2 and -4 as the germline RNAi inheritance machinery.

We asked whether, under normal reproductive conditions, the germline RNAi inheritance machinery transmits endogenous RNAi silencing signals across generations. To test this idea, we first used H3K9me3 as a read-out for endogenous nuclear RNAi in germ cells. We isolated wild-type or *nrde-2/3/4*$^{-/-}$ animals, conducted H3K9me3 chromatin immunoprecipitation (ChIP), and subjected H3K9me3 co-precipitating nucleosome core DNA to high-throughput sequencing[8]. We identified 320 predicted genes that were depleted for H3K9me3 more than twofold in both *nrde-2*$^{-/-}$ and *nrde-4*$^{-/-}$ animals relative to wild type (Fig. 3a and Supplementary Table 2). H3K9me3 ChIP, followed by directed quantitative real-time PCR (rtPCR) analysis, confirmed the *nrde-2/4* dependence of H3K9me3 at four out of four of these loci (data not shown). NRDE-dependent H3K9me3 was present in germ cells; in temperature-sensitive *glp-4(ts)* mutants[14], which lack most germ cells, H3K9me3 was significantly reduced at most *nrde-2/4*-dependent sites (Fig. 3a, $P = 2 \times 10^{-13}$). Together, these data show that *nrde-2* and *-4* contribute to H3K9me3 at several loci in germ cells. Henceforth, we refer to these loci as the endogenous NRDE germline target genes.

HRDE-1 co-precipitated with endogenous small RNAs (Fig. 3b). We sequenced these small RNAs and found that HRDE-1 bound endogenous 22G-siRNAs (22 nucleotides in length with a 5′ guanosine residue), which were expressed in germ cells, and were antisense to ~1,500 predicted coding genes (Supplementary Table 2 and Supplementary Fig. 11). HRDE-1 22G-siRNAs also targeted pseudogenes and cryptic loci (Supplementary Table 2). 22G-siRNAs are synthesized by RdRPs acting on cellular RNA templates[10,15], suggesting that the HRDE-1 22G-siRNAs are synthesized by RdRP activity in germ cells (see Supplementary Discussion). Three lines of evidence link HRDE-1 to the regulation of gene expression at NRDE germline target genes. First, we observed a correlation between genomic sites homologous to the most abundant (top 200) HRDE-1-bound siRNAs and genomic sites depleted for H3K9me3 in *nrde-2/4*$^{-/-}$ animals ($P = 2 \times 10^{-16}$) (Fig. 3c, Supplementary Fig. 12 and Supplementary Table 2). Second, we conducted H3K9me3 ChIP on *hrde-1*$^{-/-}$ animals and quantified H3K9me3 expression at fourteen NRDE germline target genes. At thirteen of these loci, H3K9me3 was depleted in *hrde-1*$^{-/-}$ animals (Fig. 3d and Supplementary Fig. 13). Third, we observed increased

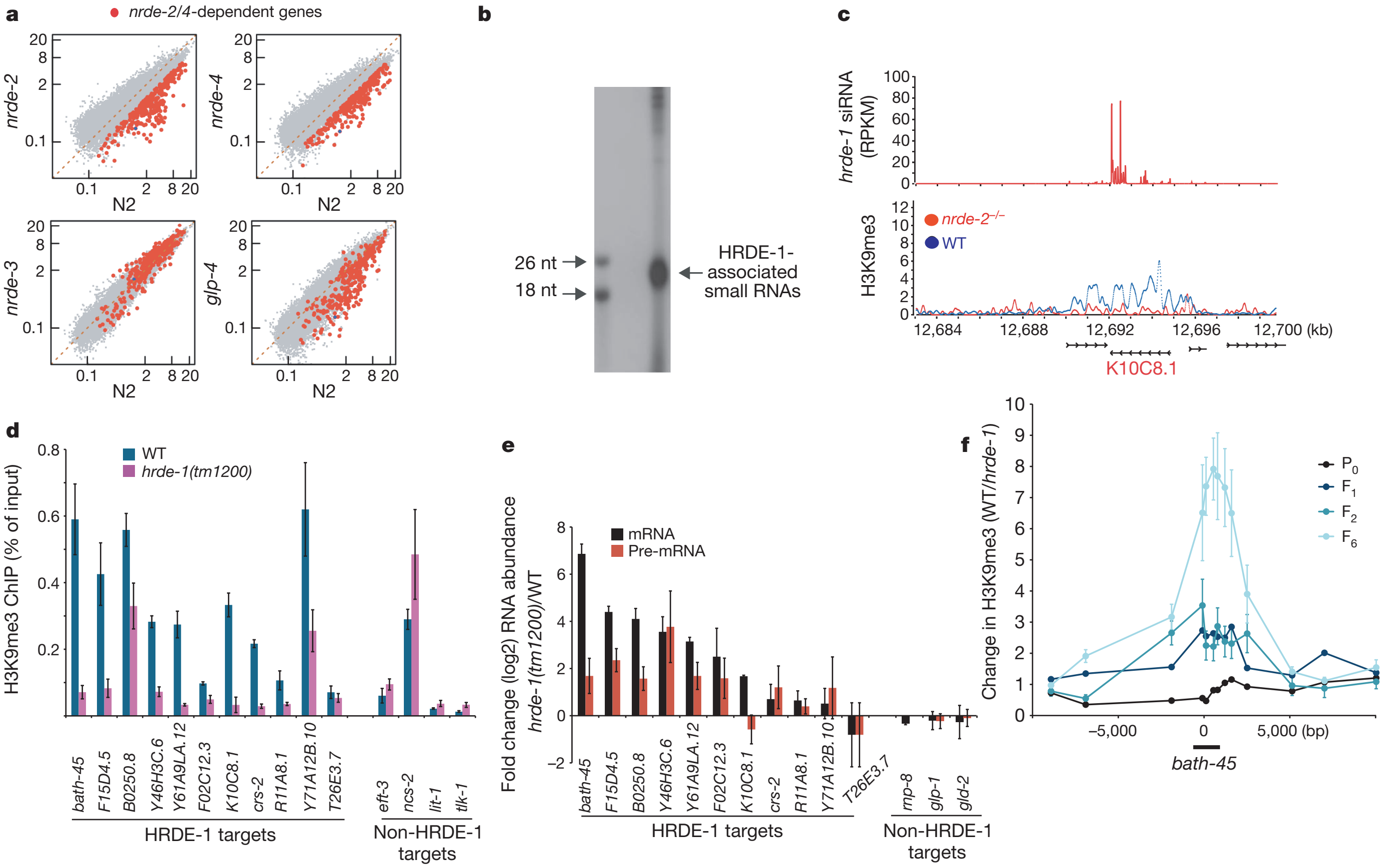


Figure 3 | The RNAi inheritance machinery transmits endogenous epigenetic information across generations. **a**, H3K9me3 in mutant (*y*-axis) versus wild type (N2, *x*-axis). Levels of H3K9me3 at *C. elegans* genes are the ratio of immunoprecipitated nucleosomes to input nucleosomes. *smg-1* (positive control)[8] (blue dot) and *nrde-2/4*-dependent genes (red) are highlighted. N2 and *nrde-2* ChIP-seq data were published previously (Gene Expression Omnibus (GEO) accession number GSE32631)[8]. **b**, Flag–HRDE-1 co-precipitating RNA was ^{32}P-radiolabelled and analysed by polyacrylamide gel electrophoresis (PAGE). nt, nucleotide. **c**, Example of an HRDE-1 target gene. RPKM, reads per kilobase (kb) per million mapped reads. **d**, rtPCR quantification of H3K9me3 ChIP (wild-type or *hrde-1*$^{-/-}$ animals grown at 25 °C) at 11 genes targeted by *hrde-1* siRNAs (HRDE-1 targets). Four genes, which are not targeted by *hrde-1* siRNAs and exhibit NRDE-independent H3K9me3 in the germ line (non-HRDE-1 targets), do not exhibit H3K9me3 loss, showing that H3K9me3 loss in *hrde-1* mutants is not simply due to loss of germ cells. Data are expressed as the percentage of input DNA recovered by ChIP; the mean ± s.d. are shown ($n = 3$). **e**, Total RNA was isolated from wild-type or *hrde-1*$^{-/-}$ animals (25 °C) and rtPCR was used to quantify mRNA or pre-mRNA levels from 11 HRDE-1-targeted genes. Three genes not targeted by HRDE-1 siRNAs, but expressed in germ cells, are also shown. Data are normalized to *nos-3* mRNA (germ line only) and expressed as a ratio (*hrde-1*$^{-/-}$/wild type); the mean ± s.d. are shown ($n = 3$). **f**, *dpy-17*$^{-/-}$ or *hrde-1*$^{-/-}$;*dpy-17*$^{-/-}$ animals were out-crossed five times, and dumpy (Dpy) adult animals (P_0) and adult progeny (F_1, F_2, F_6) were isolated (20 °C) and H3K9me3 was quantified by rtPCR. Data are expressed as a ratio (wild-type/*hrde-1*$^{-/-}$), and mean ± s.e.m. are shown ($n = 1$ for P_0 and F_1, and $n = 3$ for F_2 and F_6). Note, in this panel increased H3K9me3 signal means loss of H3K9me3 in *hrde-1*$^{-/-}$ animals.

pre-mRNA expression from many germline target genes in *hrde-1*$^{-/-}$ animals, indicating that the RNAi inheritance machinery silences germline target genes co-transcriptionally during the normal course of reproduction (Fig. 3e). These data indicate that HRDE-1 contributes to H3K9me3 in the germ line and support a model in which HRDE-1 uses endogenous 22G-siRNAs as specificity factors to direct nuclear RNAi in germ cells.

We asked whether HRDE-1-mediated nuclear RNAi at germline target genes was heritable. We out-crossed *hrde-1*$^{-/-}$ animals with wild-type animals, isolated *hrde-1*$^{-/-}$ progeny, and conducted H3K9me3 ChIP on these *hrde-1*$^{-/-}$ animals and their progeny. H3K9me3 at germline target genes was progressively lost over generations in *hrde-1*$^{-/-}$ animals (Fig. 3f and Supplementary Fig. 14). Similar results were seen with *nrde-1*$^{-/-}$ animals (Supplementary Fig. 14). Coincident with a loss of H3K9me3, germline target gene overexpression became more pronounced in late generation *hrde-1*$^{-/-}$ animals (Supplementary Fig. 15). These data show that the RNAi inheritance machinery transmits endogenous gene regulatory information across generational boundaries.

Why an organism might transmit gene regulatory information across generations is an intriguing question. During the course of our studies, we noticed that our RNAi inheritance-defective strains would periodically become sterile; stock plates would contain hundreds of adults, but no progeny. We proposed that the RNAi inheritance machinery might be required to maintain the integrity of the germ-cell lineage. To test this idea, we out-crossed two independently isolated alleles each of *hrde-1*$^{-/-}$ and *nrde-1/2/4*$^{-/-}$ to wild-type and then monitored fertility across generations. After out-crossing, *hrde-1*$^{-/-}$ and *nrde-1/2/4*$^{-/-}$ animals exhibited near wild-type fertility (early generations), but became sterile in subsequent generations (late generations) (Fig. 4a and Supplementary Fig. 16). These data show that RNAi inheritance-defective animals exhibit a mortal germline (Mrt) phenotype[16]. Animals lacking the somatic AGO NRDE-3 were not Mrt (Supplementary Fig. 17). The Mrt phenotype of *hrde-1*$^{-/-}$ animals was temperature sensitive: *hrde-1*$^{-/-}$ animals were Mrt at 25 °C, but not at 20 °C, indicating that growth at higher temperatures is required to reveal defects associated with HRDE-1 loss (Supplementary Fig. 18). Most late-generation *hrde-1*$^{-/-}$ mutants (grown at 25 °C) failed to produce mature oocytes or sperm, showing that one reason *hrde-1*$^{-/-}$ animals do not produce progeny is owing to defects in gametogenesis (Fig. 4b and Supplementary Fig. 19). Twenty-six per cent of late generation *hrde-1*$^{-/-}$ animals were able to produce sperm

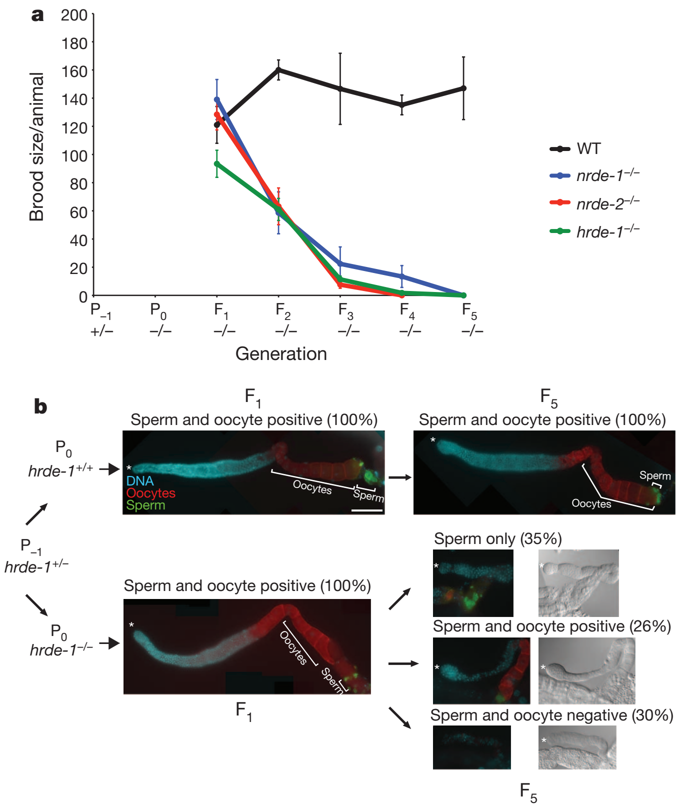


Figure 4 | The RNAi inheritance machinery promotes germline immortality. a, Animals of indicated genotypes were out-crossed to wild-type two to four times, and brood sizes scored across generations at 25 °C. Data are mean ± s.e.m. ($n = 5$). Note, for unknown reasons, *nrde-1* and *nrde-4* mutants are Mrt at both 20 °C and 25 °C, but *hrde-1* and *nrde-2* mutants are Mrt only at ~25 °C ($n = 4$). **b**, *hrde-1*$^{-/-}$ animals were out-crossed three times, and *hrde-1*$^{+/+}$ or *hrde-1*$^{-/-}$ siblings were isolated. Gonads were isolated and stained with 4′,6-diamidino-2-phenylindole (DAPI; DNA), and immunofluorescence was used to detect sperm (green) and oocytes (red) (see Methods) in F_1 and F_5 generations. Scale bar denotes 100 μm, and this is applicable to all images in **b**.

and oocytes (Fig. 4b). Most of these gametes, however, are unlikely to be functional as the fecundity of *hrde-1*$^{-/-}$ animals in this late generation was only 1% of that of wild-type animals (Fig. 4a). Finally, late generation *hrde-1*$^{-/-}$ animals exhibited a high incidence of male (Him) phenotype, suggesting that loss of RNAi inheritance may cause defects in chromosome pairing and/or segregation (Supplementary Fig. 19). We conclude that the RNAi inheritance machinery is required to maintain the immortality of the germ line and that, over generations, disabling the RNAi inheritance machinery causes progressive and diverse defects in germ-cell formation and function.

Here we show that *C. elegans* possess dedicated regulatory machinery that promotes an epigenetic memory of RNAi-silencing events that occurred in distant ancestors (Supplementary Fig. 1). The AGO HRDE-1 is at the heart of this process, binding heritably expressed specificity determinants (siRNAs) to direct nuclear RNAi and promote RNAi inheritance in germ cells. Nuclear RNAi also promotes RNAi inheritance in somatic tissues[7], indicating that nuclear gene silencing events are key determinants of RNAi memory in diverse cell types. Finally, we show that the germline RNAi inheritance machinery transmits endogenous epigenetic information across generational boundaries while promoting germline immortality (Supplementary Fig. 1). Our data suggest a model in which endogenous heritable RNAs that engage HRDE-1 act as specificity factors to direct epigenomic maintenance and immortality of the germ-cell lineage. Further work is needed to determine how defects in epigenomic maintenance relate to germline mortality (see Supplementary Discussion). We note, however, that both processes depend on the same complement of factors (*hrde-1* and *nrde-1/2/4*), and in animals lacking these factors, defects in epigenome maintenance and defects in germ-cell viability are coincident over generational time. Therefore, we propose that one biological function of the RNAi inheritance machinery is to transmit 'germline immortality' small RNAs, selected in previous generations for their ability to promote fertility, across generational boundaries to promote fertility in future generations.

Note added in proof: HRDE-1 was recently shown to act downstream of piRNAs to direct a multigenerational epigenetic memory of piRNA silencing in the germ line[19–21].

METHODS SUMMARY

RNAi. RNAi experiments were conducted as described previously[11]. The *oma-1* and *pos-1* constructs were taken from the Ahringer library.

RNA immunoprecipitation. RNA immunoprecipitations (RIPs) were performed as described previously[11], with the exception that adult animals were used for all RIPs. Adult animals were frozen and dounced 10 times before RIP. Flag–NRDE-2 protein was immunoprecipitated with an anti-Flag M2 antibody (Sigma, A2220).

ChIP. ChIP experiments were performed as described previously[12], except that gravid adult animals were used. Worms were frozen before cross-linking and were dounced 10 times before sonicating. The H3K9me3 antibody was from Upstate (07-523).

***oma-1* siRNA TaqMan assay.** The TaqMan assay was performed as described previously[7]. TaqMan probe set 1 was used in Fig. 2a (see Supplementary Methods). Unless indicated otherwise, the following mutant alleles were used in this study: *hrde-1*(*tm1200*), *nrde-1*(*gg088*), *nrde-2*(*gg091*), *nrde-3*(*gg066*) and *nrde-4*(*gg129*).

Received 14 December 2011; accepted 28 June 2012.

Published online 18 July 2012.

1. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293,** 1089–1093 (2001).
2. Jablonka, E. & Raz, G. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol.* **84,** 131–176 (2009).
3. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans. Nature* **391,** 806–811 (1998).
4. Grishok, A., Tabara, H. & Mello, C. C. Genetic requirements for inheritance of RNAi in *C. elegans. Science* **287,** 2494–2497 (2000).
5. Vastenhouw, N. L. *et al.* Gene expression: long-term gene silencing by RNAi. *Nature* **442,** 882 (2006).
6. Alcazar, R. M., Lin, R. & Fire, A. Z. Transmission dynamics of heritable silencing induced by double-stranded RNA in *Caenorhabditis elegans. Genetics* **180,** 1275–1288 (2008).
7. Burton, N. O., Burkhart, K. B. & Kennedy, S. Nuclear RNAi maintains heritable gene silencing in *Caenorhabditis elegans. Proc. Natl Acad. Sci. USA* **108,** 19683–19688 (2011).
8. Gu, S. G., Pak, J., Guang, S., Maniar, J. M., Kennedy, S. & Fire, A. Amplification of siRNA in *Caenorhabditis elegans* generates a transgenerational sequence-targeted histone H3 lysine 9 methylation footprint. *Nature Genet.* **44,** 157–164 (2012).
9. Yigit, E. *et al.* Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127,** 747–757 (2006).
10. Gu, W. *et al.* Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Mol. Cell* **36,** 231–244 (2009).
11. Guang, S. *et al.* An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* **321,** 537–541 (2008).
12. Guang, S. *et al.* Small regulatory RNAs inhibit RNA polymerase II during the elongation phase of transcription. *Nature* **465,** 1097–1101 (2010).
13. Burkhart, K. B. *et al.* A pre-mRNA-associating factor links endogenous siRNAs to chromatin regulation. *PLoS Genet.* **7,** e1002249 (2011).
14. Beanan, M. J. & Strome, S. Characterization of a germ-line proliferation mutation in *C. elegans. Development* **116,** 755–766 (1992).
15. Pak, J. & Fire, A. Distinct populations of primary and secondary effectors during RNAi in *C. elegans. Science* **315,** 241–244 (2007).
16. Ahmed, S. & Hodgkin, J. MRT-2 checkpoint protein is required for germline immortality and telomere replication in *C. elegans. Nature* **403,** 159–164 (2000).
17. Tabara, H. *et al.* The *rde-1* gene, RNA interference, and transposon silencing in *C. elegans. Cell* **99,** 123–132 (1999).
18. Tabara, H., Yigit, E., Siomi, H. & Mello, C. C. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in *C. elegans. Cell* **109,** 861–871 (2002).
19. Bagijn, M. P. *et al.* Function, targets, and evolution of *Caenorhabditis elegans* piRNAs. *Science* http://dx.doi.org/10.1126/science.1220952 (4 June 2012).
20. Ashe, A. *et al.* piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans. Cell* **150,** 88–99 (2012).

21. Shirayama, M. *et al.* piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. *Cell* **150,** 65–77 (2012).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Anderson, H. Opalicious and D. Wassarman for discussions. We thank S. Ahmed and members of the Ahmed laboratory for sharing unpublished data concerning the role of *nrde-1* in germline immortality. This work was supported by grants from the Pew and Shaw scholar's programs, and the National Institutes of Health, GM88289 (S.K.), GM37706 (A.F.) and GM069454 (J.K).

Author Contributions B.B. contributed to Figs 1a–c, 2b–d and Supplementary Figs 3, 4, 5b, 6, 8, 10 and 13. K.B. contributed to Figs 2c, 3d–f, 4a, b and Supplementary Figs 13–15, 16a–c, 17, 18 and 19c. S.G.G. and A.F. contributed to Fig. 3a–c, Supplementary Table 2 and Supplementary Figs 11 and 12. G.S. contributed to Supplementary Figs 2, 5a and 16d. A.K. and J.K contributed to Fig. 4b and Supplementary Figs 7 and 19a. H.F. contributed to Fig. 4a and Supplementary Figs 16a–c, 17, 18 and 19c. S.K. contributed to Figs 1a–d, 2a, 3b, Supplementary Table 1 and Supplementary Figs 2, 6b, 9, 10b and 19b. S.K., B.B. and K.B. wrote the manuscript.

Author Information ChIP-seq and *hrde-1* siRNA data have been submitted to the Gene Expression Omnibus (GEO) under accession number GSE38041. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.K. (sgkennedy@wisc.edu).

doi:10.1038/nature11326

Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes

Swaminathan Venkatesh[1], Michaela Smolle[1], Hua Li[1], Madelaine M. Gogol[1], Malika Saint[2], Shambhu Kumar[2], Krishnamurthy Natarajan[2] & Jerry L. Workman[1]

Set2-mediated methylation of histone H3 at Lys 36 (H3K36me) is a co-transcriptional event that is necessary for the activation of the Rpd3S histone deacetylase complex, thereby maintaining the coding region of genes in a hypoacetylated state[1,2]. In the absence of Set2, H3K36 or Rpd3S acetylated histones accumulate on open reading frames (ORFs), leading to transcription initiation from cryptic promoters within ORFs[1,3]. Although the co-transcriptional deacetylation pathway is well characterized, the factors responsible for acetylation are as yet unknown. Here we show that, in yeast, co-transcriptional acetylation is achieved in part by histone exchange over ORFs. In addition to its function of targeting and activating the Rpd3S complex, H3K36 methylation suppresses the interaction of H3 with histone chaperones, histone exchange over coding regions and the incorporation of new acetylated histones. Thus, Set2 functions both to suppress the incorporation of acetylated histones and to signal for the deacetylation of these histones in transcribed genes. By suppressing spurious cryptic transcripts from initiating within ORFs, this pathway is essential to maintain the accuracy of transcription by RNA polymerase II.

Dynamic incorporation of histones into nucleosomes over the body of genes regulates the process of gene transcription[4]. In metazoans, transcription leads to replication-independent histone exchange over ORFs, replacing histone H3 with the variant H3.3 (refs 5–8). Elongation by RNA polymerase II (RNAPII) leads to H3 exchange across ORFs of highly transcribed genes, whereas exchange is low over infrequently transcribed genes[9–12]. The ORFs of transcribed genes are enriched for the Set2-mediated H3K36me mark[3], raising the question of whether this mark regulates histone exchange over ORFs.

To measure change in histone exchange over the ORFs in a *SET2* deletion (*set2*Δ) mutant, we used the *in vivo* histone exchange yeast strain[11]. Histone exchange is measured as enrichment of induced Flag–H3, normalized to constitutive Myc–H3, in G1-arrested wild-type or *set2*Δ exchange strains by chromatin immunoprecipitation (ChIP)-on-chip (Supplementary Fig. 1). Analysis of the change in histone exchange in the *set2*Δ cells over the wild type reveals an increase over ORFs (Fig. 1a), but not at intergenic regions where Set2-mediated H3K36me is absent (Supplementary Fig. 1c). This increase was dependent on gene length, observed in ORFs longer than 1,000 bases (Fig. 1b and Supplementary Fig. 1d) and higher in less transcribed genes (Supplementary Fig. 1e) previously shown to depend on the Set2/Rpd3S pathway[3].

Both replication-dependent and replication-independent histone exchange results in the enrichment of H3 K56 acetylation (H3K56ac) on the genome[12,13]. Acetylation of H3K56 by the acetyltransferase Rtt109 occurs on soluble histones, not chromatin[14]. This acetyl mark is therefore enriched at genomic regions undergoing histone exchange. This feature was used to confirm that increased histone exchange in *set2*Δ cells occurs with endogenous H3. The genomic distribution of H3K56ac and H3 was determined in G1-arrested wild-type and *set2*Δ cells in a BY4741 background (Supplementary Fig. 2). H3K56ac increased towards the 3′ end of ORFs in *set2*Δ cells (Fig. 1c) and was dependent on gene length (Fig. 1d). Thus, *set2*Δ causes increased histone exchange and accumulation of H3K56ac over ORFs.

Loss of Set2-mediated H3K36me also increases H4 acetylation (H4ac) across ORFs[1,3]. To determine whether increased exchange correlated with increased acetylation, we examined the distribution of H4ac (normalized to Myc–H3) in the wild-type and *set2*Δ exchange strains (Supplementary Fig. 3a). In *set2*Δ cells, H4ac increased towards the 3′ end of ORFs over the wild type (Supplementary Fig. 3b) and was positively correlated with gene length (Supplementary Fig. 3c). We selected RNAPII-regulated genes that showed increased H4ac over

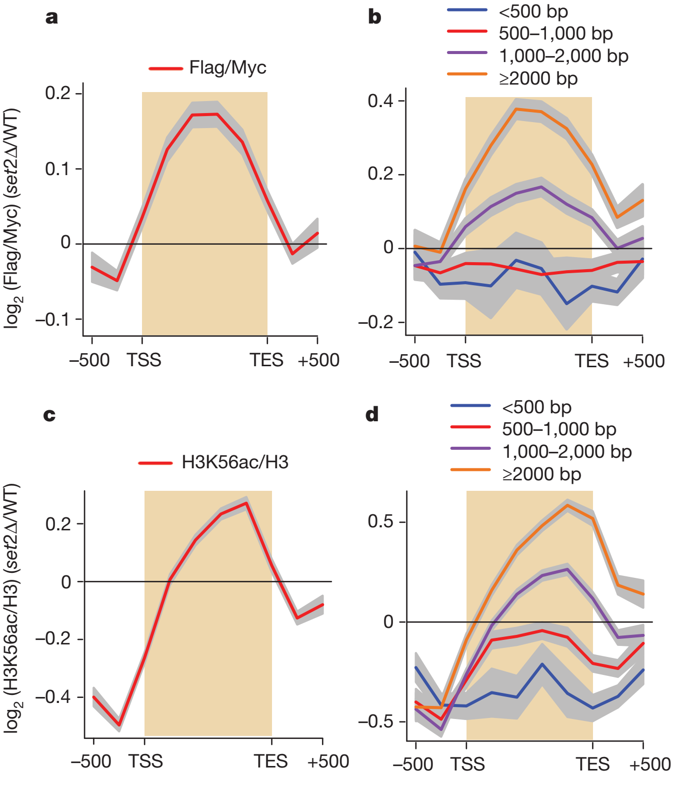


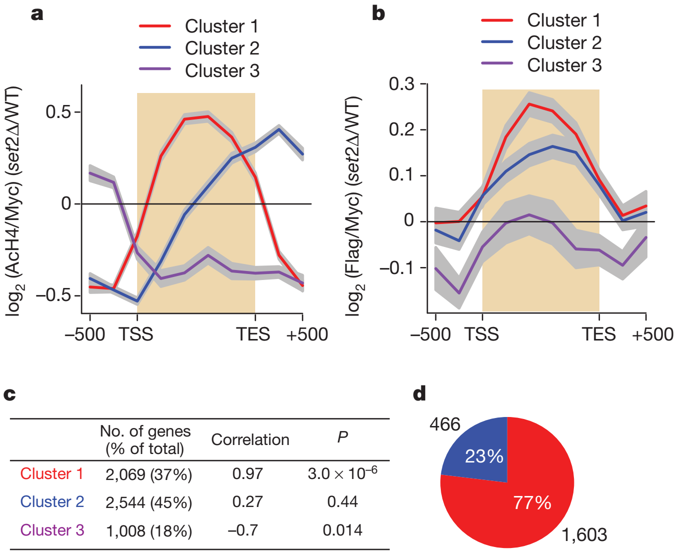
Figure 1 | Loss of Set2 results in increased histone exchange and an enrichment of H3K56ac over ORFs. Averaged $\log_2$ ratios of immunoprecipitate (IP) over input are presented as a ratio of *set2*Δ over the wild type (WT). The beige box indicates the coding region. TSS, transcription start site; TES, transcription termination site. Grey shading denotes 95% confidence interval. The keys for the distribution are shown. **a**, The genome average plot for the change in histone exchange in the *set2*Δ strain over the wild-type exchange strain. **b**, The data in **a** separated on the basis of ORF length, with the average plotted. **c**, The genome average plot for the change in H3K56ac (normalized to H3) in the *set2*Δ strain over the wild-type BY4741 strain. **d**, The data in **c** separated on the basis of ORF length, with the average plotted.

[1]Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, Missouri 64110, USA. [2]School of Life Sciences, Jawarharlal Nehru University, New Mehrauli Road, New Delhi 110067, India.

the gene (promoters and ORFs) in *set2*Δ cells over the wild type and clustered them according to the degree of similarity between their profiles of acetylation increase. We obtained three clusters of genes, the averaged profiles of which showed increased H4ac over the entire ORF (cluster 1), towards the 3′ end of ORFs (cluster 2) or at the promoter (cluster 3) (Fig. 2a). The exchange data were grouped into the same clusters as those obtained for H4ac (Fig. 2b). Averaged profiles of the first two clusters showed increased exchange across ORFs. The genes in cluster 1 showed a strong correlation between exchange and acetylation (Pearson coefficient 0.97; $P = 3.4 \times 10^{-6}$), whereas the other two clusters did not (Fig. 2c), and 77% were longer than 1,000 bases (Fig. 2d), which showed a strong correlation between acetylation and exchange (Supplementary Fig. 4a–c).

We grouped the increase profiles (Methods) of H3K56ac, H3K9ac and H4K12ac across ORFs (Supplementary Figs 5 and 6a–c) into clusters obtained for H4ac (Fig. 2a) and determined the pairwise correlation coefficients between the acetylation profiles and histone exchange for cluster 1 (Supplementary Fig. 6d, e). Each acetyl mark tested showed a strong correlation with histone exchange (Pearson coefficient more than 0.85). We observed a loss of promoter acetylation for all acetyl marks tested without affecting histone exchange (Supplementary Fig. 6d). This could be a result of deacetylase recruitment to the promoter regions[15,16].

A strong correlation indicates a causal relationship between histone exchange and acetylation across ORFs in *set2*Δ cells. Incorporation of new histones should dilute pre-existing modifications and enrich histone acetylation associated with soluble histones[17–19], as observed in *set2*Δ cells. Thus, disruption of histone exchange by deleting histone chaperone Asf1 or Rtt109 in *set2*Δ cells could lead to decreased acetylation over ORFs. Indeed, H4ac (normalized to H3) in *asf1*Δ *set2*Δ cells showed decreased levels across ORFs compared with *set2*Δ cells (Fig. 3a), whereas *asf1*Δ did not alter its distribution over ORFs (Supplementary Fig. 7a–c). We observed a similar decrease in the H4ac levels in *rtt109*Δ *set2*Δ cells compared with *set2*Δ cells (Fig. 3b and Supplementary Fig. 7d–f), thereby establishing histone exchange as a means of loading pre-acetylated histones onto ORFs on the loss of Set2-mediated H3K36me.



	No. of genes (% of total)	Correlation	*P*
Cluster 1	2,069 (37%)	0.97	3.0×10^{-6}
Cluster 2	2,544 (45%)	0.27	0.44
Cluster 3	1,008 (18%)	−0.7	0.014

Figure 2 | Increased histone exchange correlates with increased acetylation over ORFs in *set2*Δ cells. ChIP-on-chip data were plotted as in Fig. 1. **a**, The H4ac increase profiles in the *set2*Δ strain over the wild-type exchange strain were subjected to *k*-means clustering ($k = 3$) (Methods) and the average was plotted. **b**, The increase profiles of histone exchange in the *set2*Δ strain over wild-type exchange strains for the same three clusters as in **a** were averaged and plotted. **c**, Tabular representation of Pearson's correlation coefficient and the *P* value between H4ac and the histone exchange increase profiles for each cluster. **d**, Pie chart depicting the percentage of cluster 1 genes separated into ORF lengths greater than 1,000 bp (red) or less than 1,000 bp (blue). The number of genes in each group is printed outside each sector.

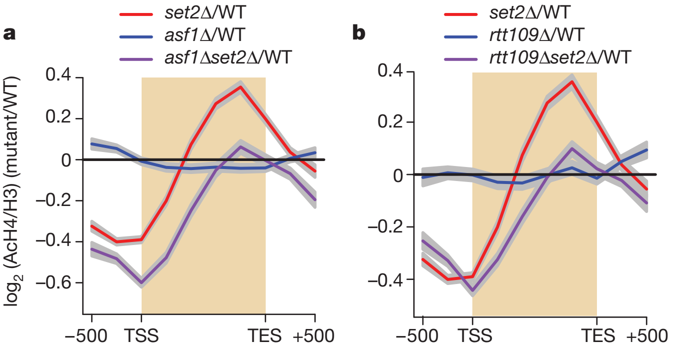


Figure 3 | Loss of Asf1-mediated exchange in *set2*Δ cells decreases the enrichment of acetylated histones over ORFs. ChIP-on-chip data were plotted as in Fig. 1. Averaged data are presented as a ratio of mutant over the wild type. **a**, Whole-genome average plots are shown for the change in acetylated H4 (normalized to H3) in *set2*Δ, *asf1*Δ and *asf1*Δ *set2*Δ strains over wild type. **b**, Whole-genome average plots are shown for the change in acetylated H4 (normalized to H3) in *set2*Δ, *rtt109*Δ and *rtt109*Δ *set2*Δ strains over wild type.

Deletion of either *SET2* or *RCO1*, a component of the Rpd3S deacetylase complex, results in the appearance of cryptic transcripts[1], implicating histone hyperacetylation over ORFs as a cause of this phenotype. Of genes that demonstrated cryptic transcription initiation from within ORFs[3], 82% also showed increased exchange in *set2*Δ cells (Fig. 4a). Consistent with data—in *asf1*Δ *set2*Δ and *rtt109*Δ *set2*Δ cells, which showed decreased histone acetylation over ORFs (Fig. 3)—cryptic transcripts from the *STE11* gene are suppressed by about 50% in these double mutants compared with the *set2*Δ cells (Fig. 4b, compare lanes 4 and 6 with lane 2, and Supplementary Fig. 8b). This is in contrast with a recent report demonstrating that *asf1*Δ *set2*Δ cells showed increased *FLO8* cryptic transcription and decreased H3K36me (ref. 20). We performed ChIP–quantitative polymerase chain reaction analysis of H3K36 trimethylation (H3 K36me3) across *PYK1* in wild-type and *asf1*Δ cells, and found that the occupancy of the mark increased over ORFs on loss of Asf1 (Supplementary Fig. 9a), as did histone H3 (Supplementary Fig. 9b).

Loss of acetylation resulting from the deletion of catalytic subunits of major lysine acetyltransferases (KAT) complexes with *set2*Δ did not significantly affect the cryptic transcript phenotype (Supplementary Fig. 8a, b). Furthermore, cryptic transcription from *STE11* remained unaffected in a histone shuffle strain with *set2*Δ and the Esa1 acetylation sites[21] mutant plasmid (H4 K5R, K8R, K12R) (Supplementary Fig. 10a, lanes 1–4, and Supplementary Fig. 10b). Whereas Rtt109 acetylates both K9 and K56 residues of H3 (ref. 17), a H3K9A mutation with *set2*Δ did not affect the cryptic transcript phenotype (Supplementary Fig. 10a, lanes 5–8, and Supplementary Fig. 10b), suggesting that Rtt109-mediated H3K56ac is crucial for co-transcriptional acetylation. However, this was untestable because the combination of H3 K56R mutation with *set2*Δ stopped cell growth (Supplementary Fig. 11a). We conclude that the loss of KAT-mediated chromatin acetylation has minimal effect on cryptic transcription regulation.

H3K36me distribution strongly anticorrelates with exchange in the wild-type strain (Supplementary Fig. 11b), and a point mutation, H3K36A, also shows cryptic transcription[1]. The combination of H3K56R with H3K36A suppressed both sense and antisense cryptic transcripts compared with the H3K36A mutant at genes known to demonstrate cryptic initiation in *set2*Δ cells (ref. 3) (Fig. 4c–e and Supplementary Figs 12–15). Loss of Set2 or H3K36me resulted in the accumulation of H3K56ac across the *STE11* ORF (Supplementary Figs 16 and 17), suggesting the occurrence of histone exchange.

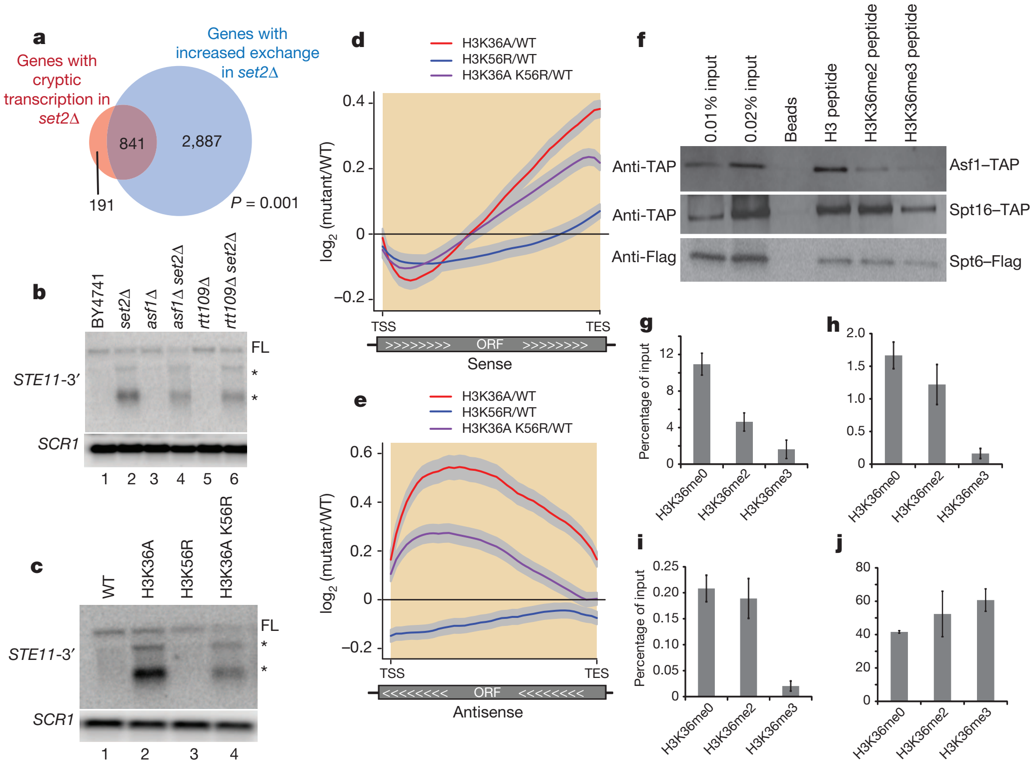


Figure 4 | Cryptic transcript phenotype caused by histone chaperone-mediated histone exchange in *set2*Δ cells. **a**, Venn diagram and hypergeometric *P* value depicting the overlap between genes showing increased histone exchange across ORFs (blue circle) and genes demonstrating cryptic initiation on loss of Set2 (ref. 3) (orange circle). **b**, **c**, Total RNA from the indicated strains were subjected to northern blotting, probing the 3′ end of the *STE11* gene (normalized to *SCR1* loading control). Full-length (FL) and cryptic transcripts (asterisk) are indicated. **b**, Northern blotting of *rtt109*Δ and *asf1*Δ deletions either singly or in combination with *set2*Δ deletion. **c**, Northern blotting of H3K56R point mutant either singly or in combination with H3K36A mutant. **d**, **e**, Messenger RNA prepared from K36A, K56R, K36AK56R and wild-type histone shuffle strains was subjected to gene expression analysis on Agilent arrays designed to detect either sense transcripts (**d**) or antisense transcripts (**e**) alone. Whole-genome gene expression ratios of mutant to wild type over ORFs of genes known to produce cryptic transcripts[3] were averaged and plotted as shown. Beige boxes indicate coding regions. Grey shading denotes 95% confidence interval. **d**, Both K56R and K36AK56R mutants show a significant decrease in sense cryptic transcription, whereas all three mutants show a decrease in full-length transcription with respect to the wild type. **e**, Both K56R and K36AK56R mutants show a significant decrease in antisense cryptic transcription. **f**, Biotinylated peptide pulldown of indicated peptides with indicated histone chaperones, analysed by western blotting. **g**–**j**, Quantification of the western blot pulldown signals with respect to the input signal and expressed as mean percentage of input (after normalizing loading amounts) for Asf1–TAP (*n* = 3) (**g**), Spt16–TAP (*n* = 2) (**h**), Spt6–Flag (*n* = 2) (**i**) and positive control Rco1–TAP (*n* = 3) (**j**). Error bars indicate s.e.m.

A H3K56Q mutant, mimicking the acetylated state in combination with H3K36A, did not affect cryptic transcript levels (Supplementary Fig. 15). We conclude that the H3K56ac-mediated histone exchange pathway regulates cryptic initiation over ORFs in *set2*Δ cells. Matching a screen identifying Asf1 as a factor that affects cryptic transcription[22], we found that 130 genes in the H3K56R mutant showed sense, but not antisense cryptic transcription, with 56% overlapping with a H3K36A mutant (Supplementary Fig. 14a, b).

Soluble histones are predominantly acetylated at H3K56 (ref. 23), H3K9 (ref. 24), and H4 K5 and K12 (ref. 25), and are enriched over ORFs in *set2*Δ cells. However, loss of H3K56ac alone causes the suppression of cryptic transcription by affecting histone exchange[13,26]. Histone exchange over the ORF therefore causes the accumulation of histone acetylation and leads to the initiation of cryptic transcription. Perturbing histone exchange by the loss of Asf1 in a *set2*Δ strain decreased acetylation but did not abolish it, establishing exchange as an important mechanism for incorporating acetylated histones on the genome in addition to targeted recruitment of KAT complexes. This is emphasized in short genes, showing increased acetylation on loss of the Set2/Rpd3S pathway without an associated increase in exchange. This effect closely correlates with the distribution of H3K36me (Supplementary Fig. 4). Set2-mediated H3K36me not only signals for deacetylation by Rpd3S but also prevents the enrichment of exchange-mediated acetylation over ORFs. To explore mechanisms of exchange inhibition over ORFs by H3K36me, we tested the interaction of histone H3 tail peptides with different histone chaperones (Asf1, Spt6 and Spt16) known to regulate histone exchange[12,27,28] and with Rco1 as a control[29]. We found that, in contrast to Rco1, interactions of the chaperones with the H3K36me3 peptide were markedly decreased (Fig. 4f–j). Moreover, whereas the H3K36 dimethylated peptide interacted with Spt6 and Spt16, its interaction with Asf1 was decreased (Fig. 4f–j and Supplementary Fig. 18). Thus, H3K36 methylation suppresses H3 interaction with histone chaperones and histone exchange over ORFs.

By suppressing histone exchange, H3K36me ensures its persistence after the transcription cycle. H3K36me can be viewed as a stable transcription mark, indicating the passage of RNAPII, which could be removed by either replication-coupled exchange or specific demethylases. Moreover, coupling histone exchange with histone acetylation is a unique method to ensure the rapid delivery of acetylation marks on the genome. This feature could help regulate global genomic events such as transcription and replication, independently of targeting KAT complexes.

METHODS SUMMARY

Data analysis. The normalized microarray data were analysed with a modified gene average analysis[3]. The ORFs of genes were divided into six equal-sized bins, and the promoter and terminator regions (500 base pairs (bp) upstream and downstream of the gene) were allocated into two bins each (Supplementary Fig. 1a). The microarray enrichment ratio ($\log_2$[immunoprecipitate (IP)/input]) for each probe was assigned to the closest bin and averaged to give rise to the bin value.

A matrix was generated with ten columns representing the coding and intergenic regions, with each row representing a gene. The columns were averaged to generate the whole-genome average plots. The enrichment of Flag and acetylated H4 in the exchange strains were normalized to Myc levels, and the enrichment of each of the modified histones was normalized to H3 levels. The ratio of the enrichment of histone exchange or histone modifications in mutant over the wild type was calculated and averaged (referred to as increase profile in the text) to generate plots represented by Fig. 1a. In these plots, the solid black line at log ratio 0 is drawn as a reference to indicate that plot lines below this reference show decreased enrichment compared to the wild type, whereas those above the reference show increased enrichment with respect to the wild type. The ORF length-dependent clusters were generated as described previously[3]. The data analysis was performed with R software.

Full Methods and any associated references are available in the online version of the paper.

Received 18 July 2011; accepted 14 June 2012.

Published online 22 August 2012.

1. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123,** 581–592 (2005).
2. Govind, C. K. *et al.* Phosphorylated Pol II CTD recruits multiple HDACs, including Rpd3C(S), for methylation-dependent deacetylation of ORF nucleosomes. *Mol. Cell* **39,** 234–246 (2010).
3. Li, B. *et al.* Infrequently transcribed long genes depend on the Set2/Rpd3S pathway for accurate transcription. *Genes Dev.* **21,** 1422–1430 (2007).
4. Workman, J. L. Nucleosome displacement in transcription. *Genes Dev.* **20,** 2009–2017 (2006).
5. Mito, Y., Henikoff, J. G. & Henikoff, S. Genome-scale profiling of histone H3.3 replacement patterns. *Nature Genet.* **37,** 1090–1097 (2005).
6. Tagami, H., Ray-Gallet, D., Almouzni, G. & Nakatani, Y. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell* **116,** 51–61 (2004).
7. Ahmad, K. & Henikoff, S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol. Cell* **9,** 1191–1200 (2002).
8. Katan-Khaykovich, Y. & Struhl, K. Splitting of H3–H4 tetramers at transcriptionally active genes undergoing dynamic histone exchange. *Proc. Natl Acad. Sci. USA* **108,** 1296–1301 (2011).
9. Kristjuhan, A. & Svejstrup, J. Q. Evidence for distinct mechanisms facilitating transcript elongation through chromatin *in vivo. EMBO J.* **23,** 4243–4252 (2004).
10. Schwabish, M. A. & Struhl, K. Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Mol. Cell. Biol.* **24,** 10111–10117 (2004).
11. Dion, M. F. *et al.* Dynamics of replication-independent histone turnover in budding yeast. *Science* **315,** 1405–1408 (2007).
12. Rufiange, A., Jacques, P. E., Bhat, W., Robert, F. & Nourani, A. Genome-wide replication-independent histone H3 exchange occurs predominantly at promoters and implicates H3 K56 acetylation and Asf1. *Mol. Cell* **27,** 393–405 (2007).
13. Kaplan, T. *et al.* Cell cycle- and chaperone-mediated regulation of H3K56ac incorporation in yeast. *PLoS Genet.* **4,** e1000270 (2008).
14. Tsubota, T. *et al.* Histone H3–K56 acetylation is catalyzed by histone chaperone-dependent complexes. *Mol. Cell* **25,** 703–712 (2007).
15. Drouin, S. *et al.* DSIF and RNA polymerase II CTD phosphorylation coordinate the recruitment of Rpd3S to actively transcribed genes. *PLoS Genet.* **6,** e1001173 (2010).
16. Kim, T. & Buratowski, S. Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5′ transcribed regions. *Cell* **137,** 259–272 (2009).
17. Fillingham, J. *et al.* Chaperone control of the activity and specificity of the histone H3 acetyltransferase Rtt109. *Mol. Cell. Biol.* **28,** 4342–4353 (2008).
18. Verreault, A., Kaufman, P. D., Kobayashi, R. & Stillman, B. Nucleosome assembly by a complex of CAF-1 and acetylated histones H3/H4. *Cell* **87,** 95–104 (1996).
19. Burgess, R. J., Zhou, H., Han, J. & Zhang, Z. A role for Gcn5 in replication-coupled nucleosome assembly. *Mol. Cell* **37,** 469–480 (2010).
20. Lin, L. J., Minard, L. V., Johnston, G. C., Singer, R. A. & Schultz, M. C. Asf1 can promote trimethylation of H3 K36 by Set2. *Mol. Cell. Biol.* **30,** 1116–1129 (2010).
21. Arnold, K. M., Lee, S. & Denu, J. M. Processing mechanism and substrate selectivity of the core NuA4 histone acetyltransferase complex. *Biochemistry* **50,** 727–737 (2011).
22. Cheung, V. *et al.* Chromatin- and transcription-related factors repress transcription from within coding regions throughout the *Saccharomyces cerevisiae* genome. *PLoS Biol.* **6,** e277 (2008).
23. Masumoto, H., Hawke, D., Kobayashi, R. & Verreault, A. A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. *Nature* **436,** 294–298 (2005).
24. Kuo, M. H. *et al.* Transcription-linked acetylation by Gcn5p of histones H3 and H4 at specific lysines. *Nature* **383,** 269–272 (1996).
25. Ai, X. & Parthun, M. R. The nuclear Hat1p/Hat2p complex: a molecular link between type B histone acetyltransferases and chromatin assembly. *Mol. Cell* **14,** 195–205 (2004).
26. Williams, S. K., Truong, D. & Tyler, J. K. Acetylation in the globular core of histone H3 on lysine-56 promotes chromatin disassembly during transcriptional activation. *Proc. Natl Acad. Sci. USA* **105,** 9000–9005 (2008).
27. Jamai, A., Puglisi, A. & Strubin, M. Histone chaperone spt16 promotes redeposition of the original H3-H4 histones evicted by elongating RNA polymerase. *Mol. Cell* **35,** 377–383 (2009).
28. Adkins, M. W. & Tyler, J. K. Transcriptional activators are dispensable for transcription in the absence of Spt6-mediated chromatin reassembly of promoter regions. *Mol. Cell* **21,** 405–416 (2006).
29. Li, B. *et al.* Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. *Science* **316,** 1050–1054 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank F. Winston and O. Rando for providing yeast strains used in this study, members of the Workman laboratory for discussion, and J. Chatfield and G. Dialynas for critical reading of our manuscript. This work was supported by National Institutes of Health grant R01GM04867 to J.L.W. and the Stowers Institute for Medical Research.

Author Contributions S.V. and J.L.W. designed the study. S.V. performed the experiments and analysed the data. M. Smolle, M. Saint and S.K. helped out with experiments. H.L. and M.M.G. wrote the R scripts used to analyse the data. K.N. provided laboratory space and resources for a part of the study. S.V. and J.L.W. wrote the manuscript.

Author Information The data discussed in this paper are deposited in NCBI's Gene Expression Omnibus under GEO accession number GSE28099 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28099). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.L.W. (jlw@stowers.org).

METHODS

Yeast strains and plasmids. The strains used in this study are listed in Supplementary Table 1. All *Saccharomyces cerevisiae* strains are derived from BY4741, S288C or w303 backgrounds. Deletions of the KAT catalytic subunits, histone chaperones and histone deacetylases were obtained from the Open Biosystems Library (maintained at the Stowers Institute Molecular Biology facility) and the deletions were confirmed by two separate PCRs with primers specific for the deletion marker and the coding region of deleted gene. *SET2* was deleted from the strain by targeted homologous integration of a PCR fragment containing the *URA3* marker flanked by the ends of the *SET2* gene. The deletions were confirmed as mentioned previously. *BAR1* was deleted in strains that were used in ChIP-on-chip experiments with a *LEU2* marker, to facilitate a stronger G1 arrest.

The histone shuffle strain was a gift from F. Winston[30]. *SET2* was deleted from the strain by homologous recombination of a PCR fragment containing the *KANMX* marker, and the deletion was confirmed by PCR. The wild-type or mutant pDM18 (ref. 30) plasmid was transformed either into the wild-type or the *SET2*-deleted strains and selected on −ura−trp double drop-out plates. The wild-type plasmid with the URA3 marker was shuffled out by selecting on −trp plates with 5-floroorotic acid (5-FOA) (1 mg ml^{-1}). The positive clones were reselected on the 5-FOA plates before being used for the northern blot analysis.

The exchange strain (MDY510) was a gift from O. Rando[11]. *SET2* was deleted from the strain by homologous recombination of a PCR fragment containing the hygromycin resistance (*HphB*) marker, and the deletion was confirmed by PCR.

The pDM18 H3 K9A and pDM18 H3 K36A plasmids were obtained from the histone mutation library generated previously[31]. The pDM18 (H4 K5, K8, K12R), pDM18 (H3 K56R), pDM18 (H3 K36A K56R), pDM18 (H3 K56Q) and pDM18 (H3 K36A K56Q) mutant plasmids were generated with a Quik Change Site-directed Mutagenesis kit by the Stowers Institute Molecular Biology facility.

Antibodies. Antibodies used in this study are listed in Supplementary Table 2.

Preparation of total RNA and mRNA. Yeast strains were grown at 30 °C in 1% yeast extract, 2% peptone, 2% dextrose (YPD) to a D_{600} of 0.8. RNA was prepared by the acid phenol extraction method as described previously[3]. The quality and quantity of total RNA produced were estimated by ultraviolet spectroscopy with a NanoDrop 2000 spectrophotometer (Thermo Scientific). Messenger RNA was purified from 3 mg of total RNA by oligo(dT)-cellulose affinity chromatography[3].

Northern blots. Northern blots were performed as described[1]. Total RNA (10 µg) was used to analyse the cryptic transcript phenotype. Probes for *STE11*, *PCA1* and *SCR1* were generated as described[32]. The blots were first probed with either *STE11* or *PCA1*, and stripped and reprobed with *SCR1* as a loading control. The blots were scanned with the Typhoon Phosphorimaging system and quantified with ImageQuant TL software (GE Healthcare). Cryptic transcripts were estimated and normalized for equal loading with the *SCR1* probe. The resultant data were either represented as the fold change over the wild-type strain or over the *set2*Δ strain. The average of three independent repeats was plotted and the standard deviation of the sample mean was calculated in each case.

ChIP assay. The exchange strains were grown in 1% yeast extract, 2% peptone (YP) plus 2% raffinose at 25 °C to a D_{600} of 0.5. α-factor was added to a final concentration of 1 µM and the cells were grown in raffinose-containing medium for a further 4 h. After the α-factor arrest, cells were collected, washed in PBS with protease inhibitors, resuspended and grown in YP + 2% galactose with 1 µM α-factor at 25 °C for 45 min. Cells were crosslinked with 1% formaldehyde at room temperature (25–28 °C) with continuous shaking for 15 min and quenched by the addition of glycine to a final concentration of 125 mM. Cells were collected by centrifugation and processed for the ChIP assay as described previously[3]. Cells were collected before and after α-factor treatment and analysed for G1 arrest by fluorescence-activated cell sorting. Cells were collected before and after galactose treatment, and cell lysates were prepared and immunoblotted with Flag and Myc antibodies to estimate the levels of tagged histone expression. The deletion of *SET2* had no effect on the expression of tagged histones. The ChIP lysates were used to immunoprecipitate Flag, Myc[11] and acetylated H4. The immune complexes were pulled down with 50 µl of Protein G Dynabeads (Invitrogen) per immunoprecipitate (IP) and processed as described previously[3]. Three biological replicates were grown for each strain and used for the ChIP-on-chip assays.

For the histone modification ChIPs, we used the deletions in strain BY4741 for analysis. This was done to ensure that the histone modification patterns observed, particularly for H3K56ac, were not an artefact resulting from histone overexpression in the exchange strain. These strains were grown in YPD to a D_{600} of 0.5. α-factor was added to a final concentration of 1 µM and the cells were grown for a further 4 h. Cells were crosslinked and processed for the ChIP assay as described previously[3]. Cells were collected before and after α-factor treatment and analysed for G1 arrest by fluorescence-activated cell sorting. The cell lysates were used to immunoprecipitate H3K56ac (2 µl per IP), H3K9ac (2 µl per IP), H4K12ac (2 µl per IP), H3 (2 µl per IP) and acetylated H4. Three biological replicates were grown for each strain and used for ChIP-on-chip assays.

ChIP–quantitative PCR assays. Quantitative PCR assays were performed as described previously[32]. DNA from ChIP assays were amplified with primers across the *STE11* (ref. 32) or *PYK1* (ref. 33) genes and normalized to the inactive *STE3* gene (forward, 5′-GTTTATGCCACCTTAGTACTGTTCGTGTTT-3′; reverse, 5′-CATATATTATAAATTGGTCTGCCAGGGTCA-3′).

Microarray analysis. *Amplification and labelling:* ChIP-on-chip assays were performed with the 4x44k yeast genome DNA arrays (Agilent, array no. 014810). The probes in this array are spaced about 290 bp apart and cover 85% of the non-repetitive sequences. Input and IP samples were used for double T7 linear amplification and labelling. In brief, up to 50 ng of IP or input DNA was treated with 2.5 U of CIP enzyme (NEB) for 1 h at 37 °C, followed by extraction with phenol–chloroform and precipitation with ethanol. The CIP-treated template was incubated with 20 U of TdT (NEB) for 20 min at 37 °C for T-tailing the ends, and the product was isolated with the MinElute Reaction Cleanup Kit (Qiagen). An anchored T7 promoter-$(dA)_{18}$ oligonucleotide was added by annealing and filling with 5 U of Exo-Klenow (NEB) for 4 h at 37 °C, followed by extraction with phenol–chloroform and precipitation with ethanol. *In vitro* transcription was performed overnight at 37 °C with the Ampliscribe T7 transcription kit (Epicentre Biotechnologies). RNAs were purified with the RNeasy Mini Kit (Qiagen) and quantified on a Nanodrop 2000 spectrophotometer (Thermo Scientific). For the second round of amplification, 100–150 ng of RNA was reverse transcribed with Superscript III reverse transcriptase (Invitrogen) and purified with the MinElute Reaction Cleanup Kit (Qiagen). The T7 promoter primer was added to the product through a Klenow filling reaction as described above, followed by *in vitro* transcription with amino allyl-UTP (Ambion) instead of UTP. Final reaction cleanup was performed with the RNeasy Mini Kit (Qiagen). For the labelling reactions, 8 µg of amino allyl-incorporated RNA in a 5 µl volume of 0.1 M carbonate buffer pH 8.7 was mixed with 5 µl (0.01 nmol) Cy3 or Cy5 dye (GE Healthcare) in dimethylsulphoxide (Sigma) and incubated at room temperature for 2 h. Inputs were labelled with Cy3 dye and IPs with Cy5 dye. Reactions were quenched with 5 µl of 4 M hydroxylamine at room temperature for 15 min and purified by RNeasy MinElute Cleanup Kit (Qiagen); dye incorporation was measured with the Nanodrop 2000.

For transcription profiling, mRNA was labelled directly with Cy dyes with the use of the Kreatech ULS system. Labelled RNA (2 µg) from wild-type, K36A, K56R and the K36AK56R strains was competitively hybridized to 4x44k arrays (Agilent nos 037746 and 037680) designed to detect transcripts specifically from either the sense or antisense strands. Ratios were calculated with the mRNA from the wild-type strain as a reference for each of the three mutants. Three biological repeats (two repeats for the K56R strain) were conducted for all microarray-based experiments.

Hybridization, data extraction and normalization: 2.5–4 µg of input and IP were combined and fragmented (Fragmentation Reagent Kit; Ambion) in accordance with the manufacturer's instructions. The hybridization mixture (Oligo CGH/ChIP-on-chip hybridization kit; Agilent) was prepared in accordance with the manufacturer's instructions with the addition of 20 µg of T7 blocking oligo[32], and hybridized overnight at 65 °C. Microarrays were washed with the Oligo aCCH/ChIP-on-chip wash buffers (Agilent) in accordance with the manufacturer's instructions. Arrays were scanned with the Agilent DNA Microarray Scanner (model no. G2505B; Agilent) and extracted with Feature Extraction software (Agilent). Data were normalized with GeneSpringGX software (Agilent). Data were read into R software and the correlation coefficient between the biological repeats was determined.

Data analysis. The normalized data were analysed with a modified gene average analysis[3]. The ORFs of genes were divided into six equal-sized bins, and the promoter and terminator regions (500 bp upstream and downstream of the gene) were allocated into two bins each (Supplementary Fig. 1a.). The microarray enrichment ratio ($\log_2$[IP/input]) for each probe was assigned to the closest bin and averaged to give rise to the bin value. A matrix was generated with ten columns representing the coding and intergenic regions, and each row representing a gene. The columns were averaged to generate the whole-genome average plots. The enrichment of Flag and acetylated H4 in the exchange strains was normalized to Myc levels, whereas the enrichment of each of the modified histones was normalized to H3 levels. The ratio of the enrichment of histone exchange or histone modifications in mutant over the wild type was calculated and averaged (increase profile) to generate plots represented by Fig. 1a. In these plots, the solid black line at log ratio 0 is drawn as a reference to indicate that plot lines below this reference show decreased enrichment compared to the wild type, whereas those above the reference show increased enrichment with respect to the wild type. The ORF length-dependent clusters were generated as described previously[3]. The data analysis was performed with R software.

For generating the gene clustering plots in Fig. 2, we first calculated the enrichment of acetylated H4 (normalized to Myc) in the *set2*Δ strain over the wild-type strain (increase profile). This ten-bin matrix was then filtered to remove genes that are not regulated by RNAPII from those genes that do not show any enrichment for acetylated H4 over the intergenic or coding regions. This removed roughly 1,300 genes, which included transfer RNA and small nuclear RNA genes and a majority of the dubious ORFs. The resultant ten-bin matrix with roughly 5,700 genes was clustered with *k*-means clustering ($k = 3$) (Partek Genomics Suite; Partek Inc.) on the basis of the correlation coefficients of the acetylated H4 increase profile in each gene. The resultant clusters were applied to the other data. The Pearson's coefficient of correlation and the *P* value was calculated with R software. The 95% confidence interval (defined as 1.96 times the s.e.m.) was calculated for all the averaged data and is represented by the grey area around the averaged trace line.

Gene expression data analysis: The normalized data were analysed by using a modified gene average analysis[3]. The ORFs of genes were divided into 40 equal-sized bins. The microarray enrichment ratio ($\log_2$[mutant/WT]) for each probe was assigned to the closest bin and averaged to give rise to the bin value. A matrix was generated with 40 columns representing the mRNA enrichment value, and each row representing a gene. The columns were averaged to generate the whole-genome average plots. The 95% confidence interval (defined as 1.96 times the s.e.m.) was calculated for all the averaged data and is represented by the grey area around the averaged trace line. For the sense transcript data, cryptic transcripts are identified by an increase in the expression ratios towards the 3′ ends of genes compared with the 5′ ends. Because cryptic transcripts initiate from within the gene body extending towards the 3′ end, the 5′ levels denote the levels of full-length transcription. In the case of the antisense cryptic transcription data, cryptic transcripts are produced towards the 5′ ends of genes, resulting in increased 5′ expression ratios.

Protein purification. TAP purifications of Asf1, Spt16 and Rco1 were performed as described previously[33]. Yeast Spt6 CDS was cloned downstream of a Flag–His tag into a baculovirus transfer vector, pBacPak8, and integrated into the baculovirus genome. Spt6–Flag protein was expressed in Sf21 cells and purified as described previously[34].

Peptide pulldown assay. The peptide pulldown assays were performed in accordance with the protocol described previously[35]. Pulldown with the unmodified or modified histone H3 peptide (22–44) was performed in buffer containing 250 mM NaCl. The beads were washed five times with the buffer.

30. Duina, A. A. & Winston, F. Analysis of a mutant histone H3 that perturbs the association of Swi/Snf with chromatin. *Mol. Cell. Biol.* **24,** 561–572 (2004).
31. Nakanishi, S. *et al.* A comprehensive library of histone mutants identifies nucleosomal residues required for H3K4 methylation. *Nature Struct. Mol. Biol.* **15,** 881–888 (2008).
32. Pattenden, S. G., Gogol, M. M. & Workman, J. L. Features of cryptic promoters and their varied reliance on bromodomain-containing factors. *PLoS ONE* **5,** e12927 (2010).
33. Mosley, A. L. *et al.* Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Mol. Cell* **34,** 168–178 (2009).
34. Hewawasam, G. *et al.* Psh1 is an E3 ubiquitin ligase that targets the centromeric histone variant Cse4. *Mol. Cell* **40,** 444–454 (2010).
35. Shi, X. *et al.* ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442,** 96–99 (2006).

doi:10.1038/nature11369

Structure of the haptoglobin–haemoglobin complex

Christian Brix Folsted Andersen[1], Morten Torvund-Jensen[1], Marianne Jensby Nielsen[1], Cristiano Luis Pinto de Oliveira[2,3], Hans-Petter Hersleth[4], Niels Højmark Andersen[5], Jan Skov Pedersen[3], Gregers Rom Andersen[6] & Søren Kragh Moestrup[1]

Red cell haemoglobin is the fundamental oxygen-transporting molecule in blood, but also a potentially tissue-damaging compound owing to its highly reactive haem groups. During intravascular haemolysis, such as in malaria and haemoglobinopathies[1], haemoglobin is released into the plasma, where it is captured by the protective acute-phase protein haptoglobin. This leads to formation of the haptoglobin–haemoglobin complex, which represents a virtually irreversible non-covalent protein–protein interaction[2]. Here we present the crystal structure of the dimeric porcine haptoglobin–haemoglobin complex determined at 2.9 Å resolution. This structure reveals that haptoglobin molecules dimerize through an unexpected β-strand swap between two complement control protein (CCP) domains, defining a new fusion CCP domain structure. The haptoglobin serine protease domain forms extensive interactions with both the α- and β-subunits of haemoglobin, explaining the tight binding between haptoglobin and haemoglobin. The haemoglobin-interacting region in the αβ dimer is highly overlapping with the interface between the two αβ dimers that constitute the native haemoglobin tetramer. Several haemoglobin residues prone to oxidative modification after exposure to haem-induced reactive oxygen species are buried in the haptoglobin–haemoglobin interface, thus showing a direct protective role of haptoglobin. The haptoglobin loop previously shown to be essential for binding of haptoglobin–haemoglobin to the macrophage scavenger receptor CD163 (ref. 3) protrudes from the surface of the distal end of the complex, adjacent to the associated haemoglobin α-subunit. Small-angle X-ray scattering measurements of human haptoglobin–haemoglobin bound to the ligand-binding fragment of CD163 confirm receptor binding in this area, and show that the rigid dimeric complex can bind two receptors. Such receptor cross-linkage may facilitate scavenging and explain the increased functional affinity of multimeric haptoglobin–haemoglobin for CD163 (ref. 4).

The release of haemoglobin (Hb) during haemolysis and tissue damage is potentially hazardous owing to the reactive properties of haem, which can engage in chemical reactions and generate free radicals[5]. After release into plasma, tetrameric Hb dissociates into αβ dimers (αβHb), and is instantly captured by the acute-phase protein haptoglobin (Hp) via a virtually irreversible interaction[6]. In humans, Hp mediates rapid clearance of Hb through high affinity-binding to the macrophage scavenger receptor CD163 (ref. 4). In addition, Hp protects tissues and cells from Hb-induced oxidative damage[7] and preserves the structural integrity of Hb through which its clearance is retained[8]. Hp–Hb complex formation also prevents renal filtration of Hb and toxic effects in the kidneys[9]. A CCP domain and a serine protease domain form the structural entities of Hp. The CCP serine protease assembly is a feature of several serine proteases, including complement factors C1s and C1r[10,11]. However, Hp is not an active protease owing to an incomplete 'catalytic triad' (Supplementary Fig. 1). Furthermore, a unique feature of Hp is the dimerization of CCP domains.

To define the structural basis for Hp-mediated recognition and protection of Hb, we determined the crystal structure of Hp–Hb purified from porcine blood to 2.9 Å resolution (Supplementary Table 1). Whereas the first crystal structure of Hb was reported more than five decades ago[12], previous attempts on structure determination of human Hp or the Hp–Hb complex have proved unsuccessful. Porcine Hp–Hb exhibits 82% sequence identity with its human counterpart (Supplementary Figs 2 and 3), and has an overall shape resembling a barbell (180 Å × 65 Å × 50 Å), with a two-fold rotational symmetry around its centre (Fig. 1a). The CCP domains connect two Hp molecules, whereas the Hp serine protease domains are responsible for Hb binding.

The Hp CCP domain (residues 33–90) has a β-sandwich arrangement similar to CCP domains in C1r/C1s[11] (Fig. 1a). CCP domains usually contain four cysteine residues forming two disulphide bridges, but the Hp CCP domain lacks a cysteine at position 68. Instead, Cys 33 engages in an interchain disulphide bridge linking two CCP domains (Fig. 1b). Our structure reveals that the two Hp CCP domains dimerize through a β-strand swap not previously observed for CCP domains (Fig. 1a and Supplementary Fig. 4). Instead of forming an anti-parallel β-sheet, strands B1 and B2 combine into a single B1/B2 strand forming an anti-parallel β-sheet with B1/B2 of the opposing CCP domain (Fig. 1b). This results in a hitherto unknown fusion CCP domain structure containing a central six-stranded β-sheet.

Electron micrographs showing Hp–Hb as a rigid structure[13] are supported by the presence of two nearly identical porcine Hp–Hb dimers in the asymmetric unit of the crystals. The rigidity is probably achieved by means of specific interactions between the CCP and serine protease domains of Hp. In particular, the side chains of Tyr 56 and Glu 135 form hydrogen bonds with each other and with the main chains of Ala 88 and Val 199 (Fig. 1c). In addition, Tyr 55, Ala 88, Val 199 and Glu 201 are involved in van der Waals contacts.

The Hp serine protease domain has the typical fold of chymotrypsin-like serine proteases, with two anti-parallel β-barrel subdomains each containing six β-strands and two or three α-helices. Several surface loop regions differ in length and conformation compared with other serine protease domains (Supplementary Fig. 1). In particular, the region designated loop 3 (residues 258–274) in serine proteases[14] is extended (Fig. 1d). This region protruding from the surface is involved in CD163 interaction because mutation of Val 259, Glu 261, Lys 262 and Thr 264 in human Hp disrupts CD163 binding[3]. These residues are conserved in porcine Hp and located at the tip of the loop. Loop 3 residues 267–271 interact directly with αHb, which may affect the conformation of the loop and influence CD163 binding.

The crystallized porcine Hp–Hb complex contains haem in oxygenated ferrous (Fe(II)) form as evidenced by the bright red colour of the crystals (Supplementary Fig. 5a), the distinct α and β absorption bands at 575 and 538 nm (Fig. 2a) and the v_4 Raman mode at 1,377 cm^{-1} (Supplementary Fig. 5b)[15]. Identical spectra are observed for porcine Hp–Hb and Hb in solution. In agreement, Hp-bound αβHb is in a similar conformation to human oxygenated αβHb[16], with

[1]Department of Biomedicine, Aarhus University, 8000 Aarhus C, Denmark. [2]Institute of Physics, University of São Paulo, Caixa Postal 66318, 05314-97, São Paulo, Brazil. [3]Department of Chemistry and iNANO Interdisciplinary Nanoscience Center, Aarhus University, 8000 Aarhus C, Denmark. [4]Department of Molecular Biosciences, University of Oslo, NO-0316 Oslo, Norway. [5]Department of Chemistry, University of Oslo, NO-0315 Oslo, Norway. [6]Department of Molecular Biology & Genetics, Aarhus University, 8000 Aarhus C, Denmark.

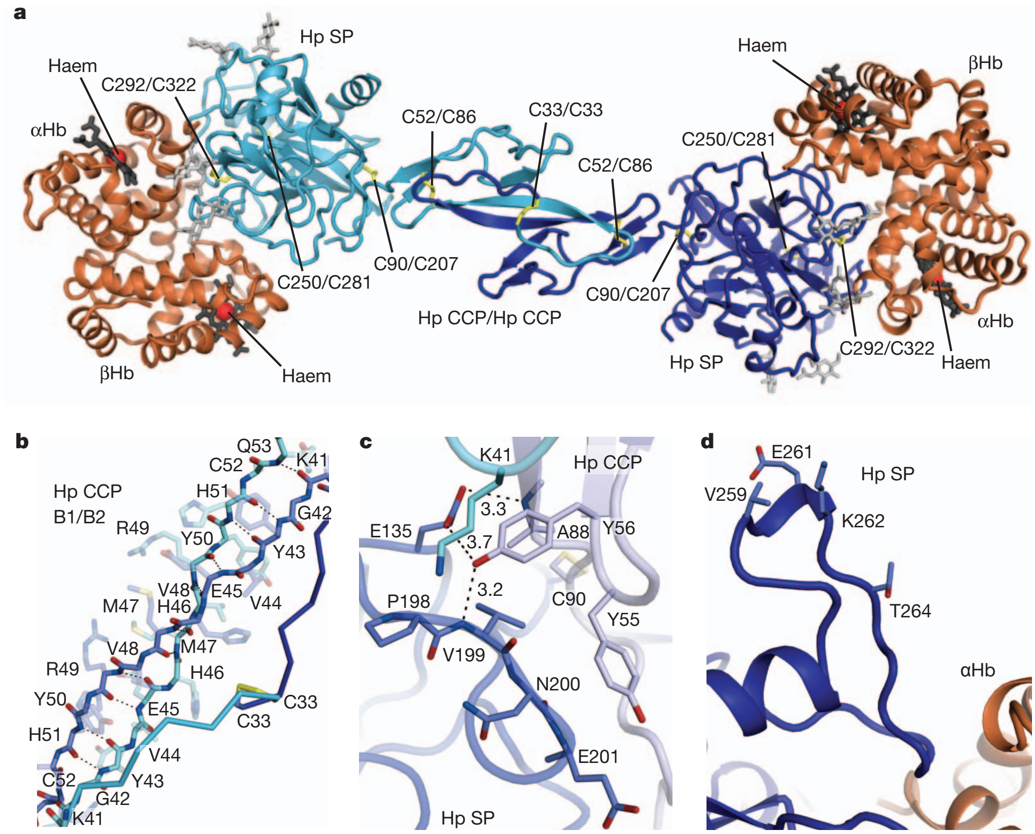


Figure 1 | Crystal structure of porcine Hp–Hb. a, Structure of porcine Hp–Hb. Hp is coloured blue and cyan, αHb and βHb are orange. Haem groups are shown as dark grey sticks. Red spheres represent Fe ions. Glycosylations are shown as light grey sticks and disulphide bridges as yellow sticks. **b**, Stick representation of the B1/B2-strands (blue and cyan) of Hp CCP domains. **c**, Interface between Hp CCP and the Hp serine protease (SP) domain. Hp SP is coloured dark blue and Hp CCP light blue. The opposing Hp CCP domain is coloured cyan. Dashed lines represent hydrogen bonds and numbers indicate bond lengths (Å). **d**, Structure of the Hp serine protease domain loop 3 region. Residues important for interaction with the CD163 scavenger receptor[3] are shown as sticks.

the Fe atom positioned in the plane of the haem group (Fig. 2b, c and Supplementary Fig. 6). This indicates that macrophages metabolise the oxygenated Hp–Hb complex, which may help to fuel the oxygen-dependent haem conversion by the haem oxygenases.

The Hb-binding site on Hp resides in surface-exposed loops of the serine protease domain, including loop A (residues 121–127), loop D (226–234), loop 1 (residues 283–289), loop 2 (residues 318–327) and loop 3 (residues 253–277) (Supplementary Fig. 1). In addition, the amino-terminal region of the Hp serine protease domain (residues 104–110) also contacts Hb. Although Hp is not an active protease, the Hb-binding site in Hp is located in the region responsible for substrate specificity in other serine proteases[14]. Furthermore, the carboxy terminus of αHb is in a position resembling the enzyme-product complex observed in C1r[17], although αHb Arg 141 is positioned outside the S1 pocket (Supplementary Fig. 7). These observations suggest that the Hp–Hb interaction originates from a product-like complex between the C terminus of αHb and an active serine protease.

Hp interacts extensively with both Hb subunits. The binding site on αHb includes residues from helix G, helix H and the FG loop, whereas residues from βHb helix C, helix G and the FG loop contact Hp (Fig. 3a). Remarkably, helix C, helix G and the FG loop also constitute the primary sites for interaction between α_1Hb and β_2Hb, and between

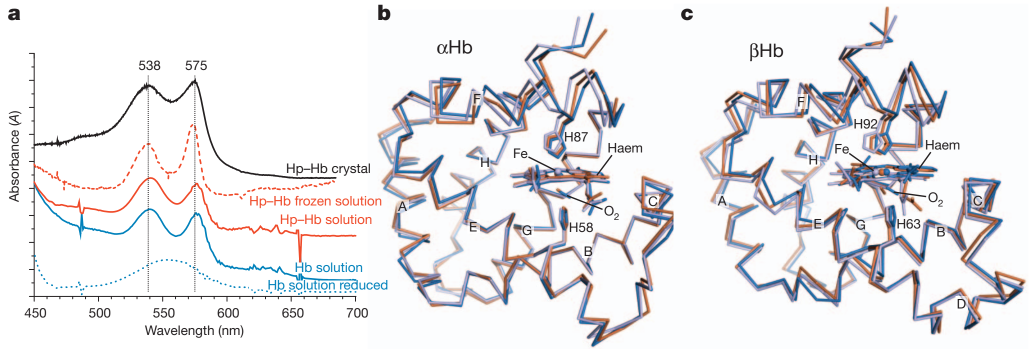


Figure 2 | The Hp-bound Hb dimer is in its oxy-state. a, Ultraviolet–visible spectra of a porcine Hp–Hb crystal (black line), porcine Hp–Hb in solution (red lines) and porcine Hb in solution (blue lines). **b**, **c**, Superimposition of porcine Hp-bound αHb (**b**) or βHb (**c**) (orange), oxygenated human αHb (**b**) or βHb (**c**) (blue, PDB accession 2DN1), and deoxygenated αHb (**b**) or βHb (**c**) (light blue, PDB accession 2DN2). Haem groups, oxygen molecules, proximal histidines (His 87 in αHb and His 92 in βHb) and distal histidines (His 58 in αHb and His 63 in βHb) are shown as sticks. Fe ions are shown as spheres.

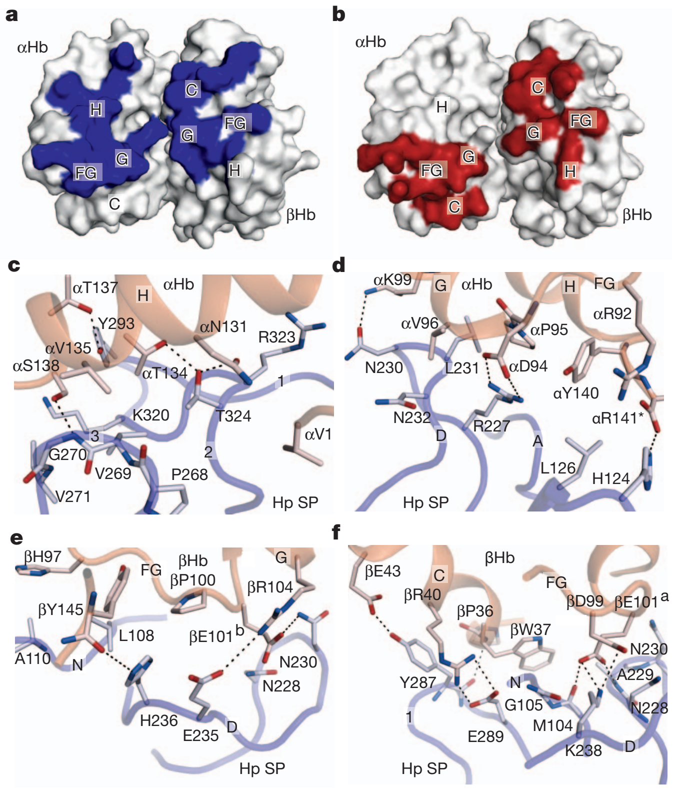


Figure 3 | The Hb contact area overlaps with the Hb dimer contact area in Hb tetramers. a, Surface representation of Hb in the Hp–Hb complex. Residues within 3.8 Å of the Hp serine protease domain are coloured blue. **b**, Surface representation of human $\alpha_1\beta_1$Hb in oxygenated tetrameric Hb (PDB accession 2DN1). α_1Hb residues within 3.8 Å of β_2Hb, and β_1Hb residues within 3.8 Å of α_2Hb, are coloured red. **c**–**f**, Selected interactions between αHb (**c**, **d**) or βHb (**e**, **f**) and the Hp serine protease domain. Residues from the Hp serine protease domain are shown as light blue sticks, and Hb residues as light orange sticks. Dashed lines represent electrostatic interactions or hydrogen bonds.

β_1Hb and α_2Hb in tetrameric Hb (Fig. 3b and Supplementary Fig. 8). This overlap explains why Hp only binds αβ dimers[2]. Furthermore, Hb tetramer formation from αβHb buries only 1,980 Å^2 compared with 2,954 Å^2 for the Hp–Hb interaction. Thus, the interaction with Hp will push the equilibrium between Hb tetramers and dimers further towards dimers. An extensive network of electrostatic interactions combined with van der Waals contacts forms the interface between Hp and Hb. Selected interactions are shown in Fig. 3c–f and a complete list of contacts in Supplementary Fig. 9. This comprehensive set of interactions fully explains the tight binding of Hp and Hb.

Tetrameric Hb is well known to undergo conformational changes that are important for regulating gas transport[18]. Several Hb residues suggested to be involved are directly recognized by Hp (for example, αVal 1, αVal 96, αLys 99, αTyr 140, αArg 141, βTrp 37 and βTyr 145; Fig. 3d–g). Hp-bound αβHb dimers exhibit non-cooperative oxygen binding with a reported P_{50} value (half-saturation oxygen tension) of 0.3 mm Hg[19]. Engineered αβHb dimers also lack cooperative oxygen binding with a reported P_{50} value of 0.59 mm Hg[20], indicating that dimeric αβHb has a high oxygen affinity compared with tetrameric Hb[18] (P_{50} value of 25–30 mm Hg in blood), irrespective of whether it is bound to Hp or not. The interactions of Hp with the FG loops and C-terminal regions of both αHb and βHb probably preserve the conformation of the F helix and consequently the position of the proximal histidine with respect to the haem group. This may explain the maintained high oxygen affinity. The high oxygen affinity indicates that deoxygenated αβHb dimers are only present under extremely low oxygen tension. Deoxygenated Hb exhibits slow binding to Hp, which has been suggested to be due to the low dissociation rate of deoxygenated Hb tetramers[21]. However, if deoxygenated αβHb retains its conformation after dissociation into αβHb dimers, several residues are not favourably positioned for interaction with Hp (Supplementary Fig. 10).

Hp protects the vascular system from damage by free Hb, but does not alter the reactive properties of the Hb haem group[8,22]. The ability of Hb to oxidize lipids and undergo structural modifications probably stems from radical intermediates formed on the globin moieties[23]. Hb residues that are specifically prone to oxidative modifications by hydrogen peroxide in the absence of Hp[8,24] are displayed on the structure of Hp–Hb in Fig. 4a. Several of these residues are located in the interface between Hp and Hb, suggesting that Hp shields the radicals formed on these residues. The Tyr 42 residue of αHb may have a key role in radical migration from α_1Hb to β_2Hb (or from α_2Hb to β_1Hb)[24]. Hp probably blocks radical migration by forcing dissociation of Hb tetramers. Other effects of hydrogen peroxide exposure are subunit dissociation, globin cross-linking and haem release[8,24,25]. These effects are probably also prevented by the tight interaction of Hp with Hb. Furthermore, Hp binds close to the haem group and it may stabilize this region of the globin moiety, which in turn may prevent haem release.

Human Hp is encoded by two different alleles (*HP1* and *HP2*) of the *HP* gene. Dimeric Hp is the gene product of the *HP1* allele, whereas the presence of the *HP2* allele gives rise to larger Hp multimers[26] (Supplementary Fig. 11). The multimerization is caused by duplication of the CCP domain[27] and the ability of each domain to dimerize with

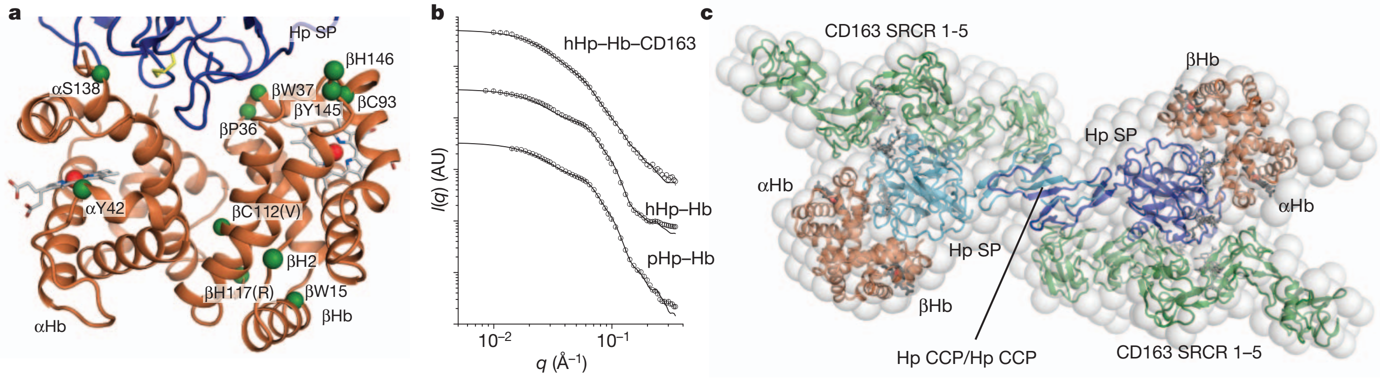


Figure 4 | Hb residues prone to oxidative modifications and SAXS. a, Human Hb residues prone to oxidative modifications indicated on the structure of porcine Hp–Hb (green spheres, corresponding porcine residues in parentheses). **b**, SAXS curves of porcine Hp–Hb (pHp–Hb), human Hp–Hb (hHp–Hb) and CD163 SRCR 1–5 bound to hHp–Hb. Circles represent experimental data and lines theoretical intensities. The structure of pHp–Hb with modelled glycosylations gives the best fit to the experimental data. AU, arbitrary units. **c**, *Ab initio* modelling of hHp–Hb bound to CD163 SRCR 1–5. The modelling on the basis of imposed P_2 symmetry (**a**, Supplementary Fig. 11) was performed using the structure of pHp–Hb with modelled glycosylations and the structure of M2BP (PDB accession 1BY2).

another CCP domain from an *HP1*- or *HP2*-encoded protein. Duplication of Cys 33 involved in CCP linkage has been the common explanation for this. However, these multimers are also probably stabilized by the formation of several CCP fusion domains (Supplementary Fig. 11). Interestingly, β-strand swapping may also occur in haptoglobin-related protein (Hpr) (Supplementary Fig. 2), which is present in Old World primates including humans, where it conveys innate immunity against trypanosome parasites[28]. Although Hpr lacks the cysteine residue at position 33 forming an interchain disulphide bridge in Hp, it associates into non-covalently linked dimers[29].

Small angle X-ray scattering (SAXS) on solutions of dimeric human and porcine Hp–Hb shows almost identical scattering curves (Fig. 4b). Furthermore, scattering curves and *ab initio* modelling (Fig. 4b, c and Supplementary Fig. 12) of dimeric human Hp–Hb incubated with a soluble recombinant CD163 fragment (SRCR domains 1–5) containing the ligand-binding site, indicate that the receptor-binding site is located in the protruding Hp loop 3 area. The data also shows that the receptor fragments can bind simultaneously to each of the Hp–Hb entities. Such receptor cross-linkage may be important for efficient CD163-mediated uptake and it explains the increased receptor avidity of the multimeric Hp–Hb complex (Supplementary Fig. 13).

METHODS SUMMARY

Hp–Hb was purified from porcine blood by a three-step chromatographic procedure (anion exchange, hydrophobic interaction and size exclusion), and crystallized using sitting-drop vapour diffusion against a reservoir containing 18% PEG3350, 10% jeffamine M-600 and 200 mM ammonium citrate, pH 7.0. Before freezing, crystals were exchanged into cryoprotection buffer containing reservoir solution with 25% PEG3350. Diffraction data were collected at 100 K at Swiss Light Source XO6SA. The structure was determined by molecular replacement, with porcine Hb (PDB accession 1QPW) and human C1r (PDB accession 2QY0) as search models. Ultraviolet–visible spectra of Hp–Hb or Hb in solution were measured on an Agilent 8453 diode array spectrophotometer and Hp–Hb crystals or frozen solution were measured on a XSPECTRA micro spectrophotometer. Raman spectra were recorded at 100 K on a Jobin-Yvon oriba T64000 instrument. Small angle X-ray scattering data were collected on a pinhole camera using a rotating anode X-ray source.

Full Methods and any associated references are available in the online version of the paper.

Received 10 May; accepted 29 June 2012.
Published online 26 August 2012.

1. Bunn, H. F., Forget, B. G. & Ranney, H. M. *Hemoglobinopathies* 308 (W. B. Saunders Company, 1977).
2. Nagel, R. L. & Gibson, Q. H. The binding of hemoglobin to haptoglobin and its relation to subunit dissociation of hemoglobin. *J. Biol. Chem.* **246,** 69–73 (1971).
3. Nielsen, M. J. *et al.* A unique loop extension in the serine protease domain of haptoglobin is essential for CD163 recognition of the haptoglobin–hemoglobin complex. *J. Biol. Chem.* **282,** 1072–1079 (2007).
4. Kristiansen, M. *et al.* Identification of the haemoglobin scavenger receptor. *Nature* **409,** 198–201 (2001).
5. Sadrzadeh, S. M., Graf, E., Panter, S. S., Hallaway, P. E. & Eaton, J. W. Hemoglobin. A biologic fenton reagent. *J. Biol. Chem.* **259,** 14354–14356 (1984).
6. Shim, B. S., Lee, T. H. & Kang, Y. S. Immunological and biochemical investigations of human serum haptoglobin: composition of haptoglobin–haemoglobin intermediate, haemoglobin-binding sites and presence of additional alleles for β-chain. *Nature* **207,** 1264–1267 (1965).
7. Lim, S. K. *et al.* Increased susceptibility in *Hp* knockout mice during acute hemolysis. *Blood* **92,** 1870–1877 (1998).
8. Buehler, P. W. *et al.* Haptoglobin preserves the CD163 hemoglobin scavenger pathway by shielding hemoglobin from peroxidative modification. *Blood* **113,** 2578–2586 (2009).
9. Fagoonee, S. *et al.* Plasma protein haptoglobin modulates renal iron loading. *Am. J. Pathol.* **166,** 973–983 (2005).
10. Gaboriaud, C., Rossi, V., Bally, I., Arlaud, G. J. & Fontecilla-Camps, J. C. Crystal structure of the catalytic domain of human complement C1s: a serine protease with a handle. *EMBO J.* **19,** 1755–1765 (2000).
11. Budayova-Spano, M. *et al.* The crystal structure of the zymogen catalytic domain of complement protease C1r reveals that a disruptive mechanical stress is required to trigger activation of the C1 complex. *EMBO J.* **21,** 231–239 (2002).
12. Perutz, M. F. *et al.* Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature* **185,** 416–422 (1960).
13. Wejman, J. C., Hovsepian, D., Wall, J. S., Hainfeld, J. F. & Greer, J. Structure of haptoglobin and the haptoglobin–hemoglobin complex by electron microscopy. *J. Mol. Biol.* **174,** 319–341 (1984).
14. Perona, J. J. & Craik, C. S. Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J. Biol. Chem.* **272,** 29987–29990 (1997).
15. Smulevich, G., Possenti, M., D'Avino, R., di Prisco, G. & Coletta, M. Spectroscopic studies of the heme active site of hemoglobin from *Chelidonichthys kumu*. *J. Raman Spectrosc.* **29,** 57–65 (1998).
16. Park, S.-Y., Yokoyama, T., Shibayama, N., Shiro, Y. & Tame, J. R. H. 1.25 Å resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. *J. Mol. Biol.* **360,** 690–701 (2006).
17. Kardos, J. *et al.* Revisiting the mechanism of the autoactivation of the complement protease C1r in the C1 complex: structure of the active catalytic region of C1r. *Mol. Immunol.* **45,** 1752–1760 (2008).
18. Dickerson, R. E. & Geis, I. *Hemoglobin: Structure, Function, Evolution and Pathology* (The Benjamin/Cummings Publishing Company, 1983).
19. Nagel, R. L., Whittenberg, J. B. & Ranney, H. M. Oxygen Equilibria of the hemoglobin–haptoglobin complex. *Biochim. Biophys. Acta* **100,** 286–289 (1965).
20. Venkatesh, B., Miyazaki, G., Imai, K., Morimoto, H. & Hori, H. Oxygen equilibrium and EPR studies on $\alpha_1\beta_1$ hemoglobin dimer. *J. Biochem.* **136,** 595–600 (2004).
21. Ip, S. H., Johnson, M. L. & Ackers, G. K. Kinetics of deoxyhemoglobin subunit dissociation determined by haptoglobin binding: estimation of the equilibrium constant from forward and reverse rates. *Biochemistry* **15,** 654–660 (1976).
22. Azarov, I. *et al.* Rate of nitric oxide scavenging by hemoglobin bound to haptoglobin. *Nitric Oxide* **18,** 296–302 (2008).
23. Miller, Y. I., Altamentova, S. M. & Shaklai, N. Oxidation of low-density lipoprotein by hemoglobin stems from a heme-initiated globin radical: antioxidant role of haptoglobin. *Biochemistry* **36,** 12189–12198 (1997).
24. Pimenova, T. *et al.* Quantitative mass spectrometry defines an oxidative hotspot in hemoglobin that is specifically protected by haptoglobin. *J. Proteome Res.* **9,** 4061–4070 (2010).
25. Jia, Y., Buehler, P. W., Boykins, R. A., Venable, R. M. & Alayash, A. I. Structural basis of peroxide-mediated changes in human hemoglobin: a novel oxidative pathway. *J. Biol. Chem.* **282,** 4894–4907 (2007).
26. Smithies, O. & Walker, N. F. Notation for serum-protein groups and the genes controlling their inheritance. *Nature* **178,** 694–695 (1956).
27. Maeda, N., Yang, F., Barnett, D. R., Bowman, B. H. & Smithies, O. Duplication within the haptoglobin *Hp2* gene. *Nature* **309,** 131–135 (1984).
28. Vanhollebeke, B. *et al.* A haptoglobin–hemoglobin receptor conveys innate immunity to *Trypanosoma brucei* in humans. *Science* **320,** 677–681 (2008).
29. Nielsen, M. J. *et al.* Haptoglobin-related protein is a high-affinity hemoglobin-binding plasma protein. *Blood* **108,** 2846–2849 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to G. Ratz for technical assistance, R. Kidmose, N. S. Laursen and the staff at the Swiss Light Source beamlines for help with data collection. We thank G. V. Jensen for performing SAXS measurements and R. E. Weber for scientific discussion. The research was supported by The Lundbeck Foundation, The Novo Nordisk Foundation, The Research Council of Norway, The European Research Council and The Danish Council for Independent Research.

Author Contributions C.B.F.A.: purification, crystallization, data collection, structure determination and analysis, manuscript preparation and study design. M.T.-J.: purification, crystallization, data collection and structure determination. M.J.N.: cloning, expression and purification. C.L.P.d.O.: SAXS measurements and analysis. H.-P.H.: ultraviolet–visible spectroscopy and analysis. N.H.A.: Raman spectroscopy and analysis. J.S.P.: SAXS measurements and analysis. G.R.A.: study design. S.K.M.: manuscript preparation and study design.

Author Information Atomic structure factors and coordinates have been deposited at the Protein Data Bank under accession number 4F4O. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.B.F.A. (cbfa@biokemi.au.dk) or S.K.M. (skm@biokemi.au.dk).

METHODS

Purification of Hp–Hb from porcine blood. Anti-coagulant (trisodium citrate and EDTA) was added to fresh porcine blood to a final concentration of 15 mM (trisodium citrate) and 0.15 mM (EDTA). Plasma and blood cells were separated by centrifugation at 4,000*g* for 20 min. Clotting factors were removed from the plasma fraction by the addition of 25 mM $BaCl_2$, followed by incubation on ice overnight and centrifugation at 8,000*g* for 15 min. The blood cell fraction was lysed by the addition of water (1:1 ratio), and cell debris was removed by centrifugation at 8,000*g* for 15 min. Serum and blood cell fractions were stored at −80 °C.

Thawed serum and blood cell fractions were mixed in a ratio of 25:1 and incubated at 4 °C overnight. The sample was diluted 1:5 in 20 mM acetic acid, pH 5.3, 10% glycerol, and loaded on a Q Sepharose Fast Flow column (GE Healthcare) equilibrated in buffer Q-A (20 mM KCl, 20 mM acetic acid, pH 5.3, 10% glycerol). A gradient from 5 to 55% buffer Q-B (500 mM KCl, 20 mM acetic acid, pH 5.3, 10% glycerol) was applied. The pH of the eluted fractions was adjusted by the addition of Tris-HCl, pH 7.6, to a final concentration of 50 mM.

Source Q fractions containing Hp–Hb were pooled and ammonium sulphate added to 60% saturation. The sample was centrifuged at 27,000*g* for 20 min and the supernatant was loaded on a Source 15 Iso column (GE Healthcare) equilibrated in buffer Iso-A (60% ammonium sulphate, 20 mM Tris-HCl, pH 7.6). A gradient from 0 to 60% buffer Iso-B (20 mM Tris-HCl, pH 7.6) was applied and Hp–Hb-containing fractions were pooled and concentrated using an Amicon Ultra centrifugal filter (10 kDa molecular mass cut-off, Millipore).

The sample was further purified using a Superdex 200 column (GE Healthcare) equilibrated in 75 mM KCl, 20 mM Tris-HCl, pH 7.6, 0.5 mM EDTA. Fractions containing >98% pure Hp–Hb were pooled and concentrated to 10 mg ml^{-1} using a Vivaspin 500 centrifugal filter (10 kDa molecular mass cut-off, GE Healthcare).

Crystallization and data collection. Crystals were obtained at 4 °C using the sitting-drop vapour diffusion method by mixing 2 μl protein solution (10 mg ml^{-1}) with 2 μl reservoir solution containing 18% PEG 3350, 10% jeffamine M-600 and 200 mM ammonium citrate, pH 7.0. Before flash-freezing in liquid nitrogen, crystals were exchanged into cryo-protection buffer containing 25% PEG 3350, 10% jeffamine M-600 and 200 mM ammonium citrate, pH 7.0. X-ray data were collected at the XO6SA beamline (Swiss Light Source) using a wavelength of 1.0 Å and at a temperature of 100 K.

Structure determination. Data were indexed, integrated and scaled with the XDS-package[30]. Initial phases were calculated by molecular replacement using the program PHASER[31]. The structures of porcine Hb (PDB accession 1QPW) and human C1r (PDB accession 2QY0) were used as search models. The model was refined using iterative cycles of refinement in PHENIX[32] followed by model building using the program 'O'[33]. Both B-factors and atomic coordinates were restrained by tight four-fold (Hb and Hp serine protease domain) or two-fold (Hp CCP) non-crystallographic symmetry with one B-factor group per residue throughout the refinement procedure. PISA[34] and PROCHECK[35] were used for structure analysis and validation. The Ramanchandran plot statistics shows 90.7% residues in the most favoured region, 9.0% in the additionally allowed region and 0.3% in the generously allowed region. Figures were prepared with PyMOL[36] and ALINE[37].

Ultraviolet–visible and Raman spectroscopy. The ultraviolet–visible spectra of porcine Hb and Hp–Hb in solution at room temperature were measured in a 1-cm quartz cuvette on an Agilent 8453 diode array ultraviolet–visible spectrophotometer. The ultraviolet–visible spectra of porcine Hp–Hb as frozen solution and as crystals were measured using a microspectrophotometer model XSPECTRA (4DX System AB) equipped with a halogen lamp and iDUS CCD detector with a Shamrock monochromator (Andor Technology). These measurements were performed at 100 K in a nitrogen cold stream (Oxford Cryosystems). The frozen solution sample was generated by mixing the Hp–Hb solution with ~50% glycerol. The spectra of Hp–Hb, both as frozen solution and as crystals, were recorded with the sample placed in a nylon loop (Hampton Research). Raman spectra were recorded at 100 K on a Jobyn Yvon Horiba T64000 instrument with the laser generated through a Millennia Pro 12sJS Nd:YVO_4 a Matisse TR ring laser and a Wavetrain frequency doubler.

SAXS. SAXS data were collected on a pinhole camera using a rotating anode as X-ray source and Göbel mirrors as optics[38]. Human Hp (phenotype Hp1-Hp1) was purchased from Sigma, and CD163 SRCR 1–5 was expressed and purified as described previously[39]. Samples were measured at 20 °C in reusable quartz capillaries with a diameter of 1.5 mm. The data are displayed as a function of the modulus of the scattering vector, $q = (4\pi/\lambda)\ \sin(\theta)$, in which $\lambda = 1.54$ Å is the X-ray wavelength and 2θ is the angle between the incident and scattered X-rays. Background subtraction and normalizations were made using the SUPERSAXS package (C.L.P.d.O. and J.S.P., unpublished). The data were normalized to absolute scale using a pure water sample as a primary standard. In all cases, a concentration series (from 2 to 8 mg ml^{-1}) was investigated to check for concentration effects. The initial analysis of the data was performed using the indirect Fourier transformation procedure[40–42]. *Ab initio* structure determination was performed using the programs DAMMIN[43] and GASBOR[44]. Modelling using known atomic resolution structures was done using the programs CRYSOL[45], BUNCH, SASREF and CORAL[46]. For the set of generated models both the average model and the most probable model were determined using the program DAMAVER[47].

30. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.* **26,** 795–800 (1993).
31. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
32. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
33. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
34. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).
35. Laskowski, R., MacArthur, M., Moss, D. & Thornton, J. Procheck – a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).
36. The PyMOL Molecular Graphics System v. 1.3r1 (Schrödinger, LLC, 2010).
37. Bond, C. S. & Schuettelkopf, A. W. ALINE: a WYSIWYG protein-sequence alignment editor for publication-quality alignments. *Acta Crystallogr. D* **65,** 510–512 (2009).
38. Pedersen, J. A flux- and background-optimized version of the NanoSTAR small-angle X-ray scattering camera for solution scattering. *J. Appl. Cryst.* **37,** 369–380 (2004).
39. Madsen, M. *et al.* Molecular characterization of the haptoglobin·hemoglobin receptor CD163. Ligand binding properties of the scavenger receptor cysteine-rich domain region. *J. Biol. Chem.* **279,** 51561–51567 (2004).
40. Oliveira, C. L. P. *et al.* A SAXS study of glucagon fibrillation. *J. Mol. Biol.* **387,** 147–161 (2009).
41. Pedersen, J. S., Hansen, S. & Bauer, R. The aggregation behavior of zinc-free insulin studied by small-angle neutron scattering. *Eur. Biophys. J.* **22,** 379–389 (1994).
42. Glatter, O. New method for evaluation of small-angle scattering data. *J. Appl. Crystallogr.* **10,** 415–421 (1977).
43. Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76,** 2879–2886 (1999).
44. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80,** 2946–2953 (2001).
45. Svergun, D., Barberato, C. & Koch, M. CRYSOL — A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28,** 768–773 (1995).
46. Petoukhov, M. V. & Svergun, D. I. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* **89,** 1237–1250 (2005).
47. Volkov, V. & Svergun, D. Uniqueness of *ab initio* shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36,** 860–864 (2003).

CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature11422

Corrigendum: 'Big Bang' tomography as a new route to atomic-resolution electron tomography

Dirk Van Dyck, Joerg R. Jinschek & Fu-Rong Chen

Nature **486,** 243–246 (2012); doi:10.1038/nature11074

In this Letter, the name of Joerg R. Jinschek (FEI Europe, Europe NanoPort, Achtseweg Noord 5, 5651 CG Eindhoven, The Netherlands) should be included in the author list. On page 244, this sentence should be deleted: "The experimental data were obtained from ref. 11.". The Author Contributions section should include the sentence: "J.R.J. provided the experimental images of graphene.". The Acknowledgements section should include the sentence: "J.R.J. thanks E. Yucelen, R. Dunin-Borkowski and Ch. Kisielowski for support, and N. Alem and A. Zettl for the gift of the graphene sample.". The PDF and HTML versions of the original paper have been corrected online.

ERRATUM

doi:10.1038/nature11426

Erratum: Fractal morphology, imaging and mass spectrometry of single aerosol particles in flight

N. D. Loh, C. Y. Hampton, A. V. Martin, D. Starodub, R. G. Sierra, A. Barty, A. Aquila, J. Schulz, L. Lomb, J. Steinbrener, R. L. Shoeman, S. Kassemeyer, C. Bostedt, J. Bozek, S. W. Epp, B. Erk, R. Hartmann, D. Rolles, A. Rudenko, B. Rudek, L. Foucar, N. Kimmel, G. Weidenspointner, G. Hauser, P.Holl, E. Pedersoli, M. Liang, M. S. Hunter, L. Gumprecht, N. Coppola, C. Wunderer, H. Graafsma, F. R. N. C. Maia, T. Ekeberg, M. Hantke, H. Fleckenstein, H. Hirsemann, K. Nass, T. A. White, H. J. Tobias, G. R. Farquar, W. H. Benner, S. P. Hau-Riege, C. Reich, A. Hartmann, H. Soltau, S. Marchesini, S. Bajt, M. Barthelmess, P. Bucksbaum, K. O. Hodgson, L. Strüder, J. Ullrich, M. Frank, I. Schlichting, H. N. Chapman & M. J. Bogan

Nature **486,** 513–517 (2012); doi:10.1038/nature11222

In this Letter, author M. S. Hunter was incorrectly listed as M. M. Hunter; this has been corrected online in the PDF and HTML of the original paper.

ERRATUM

doi:10.1038/nature11454

Erratum: Structured spheres generated by an in-fibre fluid instability

Joshua J. Kaufman, Guangming Tao, Soroush Shabahang, Esmaeil-Hooman Banaei, Daosheng S. Deng, Xiangdong Liang, Steven G. Johnson, Yoel Fink & Ayman F. Abouraddy

Nature **487,** 463–467 (2012); doi:10.1038/nature11215

In this Letter, the received date was incorrectly listed as 20 December 2012 instead of 20 December 2011; this has been corrected in the HTML and PDF versions of the manuscript.

ERRATUM

doi:10.1038/nature11455

Erratum: Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy

N. M. Archin, A. L. Liberty, A. D. Kashuba, S. K. Choudhary, J. D. Kuruc, A. M. Crooks, D. C. Parker, E. M. Anderson, M. F. Kearney, M. C. Strain, D. D. Richman, M. G. Hudgens, R. J. Bosch, J. M. Coffin, J. J. Eron, D. J. Hazuda & D. M. Margolis

Nature **487,** 482–485 (2012); doi:10.1038/nature11286

In the print and PDF versions of this Letter, the statement at the end of the paper that there was Supplementary Information linked to the online version of the paper was incorrect; this statement has now been removed in the PDF version of the manuscript.

CAREERS

TURNING POINT Seismologist's open-minded approach brings recognition **p.463**

@NATUREJOBS Follow us on Twitter for the latest news and features **go.nature.com/e492gf**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

B. MELLOR

POSTDOCS

A voice for the voiceless

In its first ten years, the US National Postdoctoral Association has helped to raise the profile of postdocs. But championing their cause still presents challenges.

BY KAREN KAPLAN

Before Alyson Reed became head of the US National Postdoctoral Association (NPA), she had only the vaguest ideas about what a postdoctoral researcher does.

Reed was hardly alone. After she took the job as the NPA's inaugural executive director in 2003, she learned that few outside science and academia knew what postdocs are or do. Even on university campuses, many postdocs felt invisible and anonymous, crucial to research but suspended in limbo with no means of networking, creating a community or being heard.

Almost ten years of efforts by the NPA have helped to mitigate those challenges. Based in Washington DC, the non-profit organization has worked hard on behalf of its 2,700 members and the nation's more than 60,000 postdocs (a number based on reports from member institutions). It has helped stakeholders — including federal agencies, members of Congress and policy-makers — to become eminently familiar with what postdocs are, what they do and the conditions they face. It has raised the issue of shoddy compensation and highlighted the difficulties of career development.

Yet most US academic postdocs still work long hours for trifling pay and have no clear route into a permanent position. Observers say that the NPA has made progress, but should do more. The association would like to boost outreach and advocacy and offer more services. But a meagre budget, a small staff and funding challenges present significant obstacles.

GRASS-ROOTS BEGINNING

The seeds of the NPA were sown in 2002, during a postdoc networking meeting launched and supported by the American Association for the Advancement of Science (AAAS) in Washington DC. Two years earlier, a report[1] by the US National Academies had recommended that funding agencies, institutions and other bodies award postdocs appropriate recognition and compensation; provide health benefits; and ease the transition to a permanent position. But few had adopted the proposals by the time of the AAAS meeting.

So seven postdocs at the meeting decided to change the landscape themselves. "We felt it was the right time to try to have conversations with institutions and funding agencies about the issues — career development and appropriate salary and benefits — and to try to come up with federal policies to support them," says Carol Manahan, one of the founding members of the NPA and now associate director of the education office at Novartis Institutes for BioMedical Research in Cambridge, Massachusetts.

"We were ghosts," says Orfeu Buxton, another founder and now a neuroscientist at Brigham and Women's Hospital in Boston, ▶

Massachusetts. “We formed the NPA in part because we wanted to bring a face to the faceless, a voice to the voiceless.” They resolved to establish a group that could represent and advocate for postdocs around the country.

Over the course of a few months, the group created a steering committee and partnered with the AAAS, which provided office space, a telephone line and advice on how to network and correspond with heads of federal agencies and non-profit groups, and how to raise and manage money. The NPA pursued the Alfred P. Sloan Foundation in New York for funding, and won a US$10,860 planning grant, followed by a $456,000, 18-month operating grant, which the AAAS received and administered.

The organization had no track record or reliable sources of future funding, but it was still worth investing in, recalls Michael Teitelbaum, a senior adviser at Sloan. “It was a high-risk venture, but the proposal was really good.”

With office space and funding in hand, the NPA held its first annual meeting in March 2003 in Berkeley, California, drawing some 80 postdoc attendees from various universities. Six months later, Reed came on board. “The system was broken,” she recalls learning during her first few months on the job. “Postdocs needed a seat at the table with decision-makers to set standards, create infrastructure and improve outcomes.”

GETTING THE WORD OUT

Before the NPA could advocate for postdocs — not to mention collect data about their roles at research institutions — it had to define what a postdoctoral researcher is, says Reed. The association helped the US National Institutes of Health (NIH) and the US National Science Foundation (NSF) to adopt a formal definition in 2007. That definition — which states, in part, that postdocs are “engaged in a temporary and defined period of mentored advanced training to enhance the professional skills and research independence” — helps institutions and principal investigators to see postdocs as trainees and protégés seeking to advance their careers, rather than as just a pair of hands at the bench, says Reed.

“The system was broken. Postdocs needed a seat at the table with decision-makers.”

That done, the association staked out its position. As a 501(c)(3) non-profit organization, the NPA is allowed to spend up to 20% of its resources on lobbying — trying to influence law-makers, including Congress. But current executive director Cathee Johnson Phillips says that the organization is better suited to advocating, recommending and educating. In 2009, it created National Postdoc Appreciation Day, since extended to a week and now observed at about 90 US and Canadian institutions. In 2010, the US House of Representatives officially recognized the week and the contribution of postdocs to the scientific enterprise. That, says Johnson Phillips, “was the first step in educating the legislative branch of our government regarding the importance of the postdoc in US research and discovery”.

Manahan represented the NPA on a US National Academies committee that called on[2] the NIH to provide better support for postdocs seeking to set up their own labs. The committee’s report helped to persuade the NIH to establish its Pathway to Independence grants in 2006. The NPA’s input, including policy papers and talks with congressional subcommittee staff, helped to prompt 2007 legislation that required all NSF grant proposals that include funds for postdocs to list mentoring activities, and all NIH proposals that include postdoc funds to expand postdoc data collection. And last year, the NPA submitted a 19-page response to the NIH’s request for information for a working group on the biomedical workforce. This June, the working group released the final draft of its report[3] on what the workforce is lacking. “Their input certainly helped our analysis and identified needs that had not been fully realized,” says NIH director Francis Collins, who was the keynote speaker at the NPA’s 2010 annual meeting and remains sympathetic to the association. “The recommendations they put forward are under serious consideration.”

In 2003, the NPA teamed up with Sigma Xi, a researchers’ society based in Research Triangle Park, North Carolina, and others to conduct a multi-campus survey of postdocs — the first collection of data on postdocs’ work, career goals and perceptions of policies and practices at their institutions. Among other results, it found[4] that postdocs with structured plans for oversight and career development had fewer conflicts with their advisers, rated their advisers more highly and had more publications than those without such plans.

But perhaps the NPA’s biggest accomplishment has been in encouraging US universities to set up their own on-campus postdoctoral offices and associations, for which it provides online toolkits (see go.nature.com/njxrwo) and conducts site visits. The NPA’s annual meeting provides a forum in which such offices and associations can share ideas and best practices.

The association now has about 130 member offices on US university and other research campuses, and has inspired the creation of postdoc organizations in other countries (see ‘International influence’). The offices support postdocs, offering networking and social events, handbooks and help with grant writing and presentations. They also often serve as the first line of action for grievances. A postdoc-office administrator can bring together faculty members and people from human-resources departments to create an institutional culture that improves conditions for postdocs.

A DOLLAR SHORT

Still, with an annual budget of just $600,000 from membership dues, grants and marketing, and just four staff members — one of them part time — including Johnson Phillips, the organization struggles to reach a wider audience. Johnson Phillips has a long list of projects that she would pursue if funding were more plentiful: a follow-up to the Sigma Xi survey; a smartphone app; more travel awards to bring administrators and faculty members to NPA meetings; more marketing; more staff members to develop white papers and other educational tools; and exhibits at many more conferences. She would

GLOBAL ACTION

International influence

The US National Postdoctoral Association (NPA) has extended its influence abroad, with several fledgling postdoc organizations looking to it for advice. They hail from Canada, the United Kingdom, Ireland, France, China, Japan and Qatar, says Cathee Johnson Phillips, executive director of the NPA in Washington DC.

The Canadian Association of Postdoctoral Scholars (CAPS-ACSP), for example, formed in 2007. Canadian postdocs can join the NPA as affiliate members, but some wanted to form their own group to tackle issues in Canada. CAPS-ACSP received mentoring and guidance from the NPA, as well as initial financial aid to set up teleconferences for its steering committee. The association maintains a website and Facebook page on which it posts news and issues that affect postdocs; it has also conducted a survey of postdocs, released white papers and presented talks.

The UK Research Staff Association (UKRSA) formed in 2006 and is funded by Vitae, an independent organization based in Cambridge that receives most of its money from the UK government’s research councils. “The NPA was a big inspiration,” says Daniel Weekes, co-chair of the UKRSA. “Their success certainly spurred us on.”

The NPA is also one of the founders of the International Consortium of Research Staff Associations, a global organization established in July that aims to support researchers around the world and give them a voice. “People from other countries want to talk with us because they know of the NPA and what it does,” says Johnson Phillips, “and they want to learn from us about how to support such a community.” **K.K.**

also like, to the extent allowed for a non-profit, to hold awareness-raising meetings with federal law-makers or their staff.

Some members say that the NPA should redouble its efforts to regularly collect and analyse data on postdocs. Johnson Phillips says that the association doesn't have the funding for such endeavours, although she notes that in 2009, it conducted a pilot poll of postdocs on issues including compensation, benefits and career pathways as a follow-up to the Sigma Xi project. Other observers say that the NPA needs to increase its advocacy work. John Scatizzi, an immunology postdoc and president of the postdoc association at the Scripps Research Institute in La Jolla, California, would like to see greater progress on standardizing compensation and benefits. "They need to take more of an active role," he says.

To further improve postdoc career prospects, says Collins, the NPA should boost its visibility in the academic community. "This is a great opportunity for them to align themselves with organizations such as the Association of American Universities and the Association of Public and Land-grant Universities," says Collins. "These relationships could help them to further their agenda. A lot of what postdocs need in terms of career growth is controlled not by the NIH but by universities."

Johnson Phillips is quick to point to ongoing advocacy and educational efforts at the NPA. The association regularly publishes white papers on postdoc issues, responds to federal agencies' requests for information and makes recommendations to the agencies, NPA member institutions and other stakeholders. With the Association of Public and Land-grant Universities in Washington DC, the NPA is developing a national certification programme to identify institutions that follow its best practices and recommendations. It has also developed a set of core postdoc competencies for evaluating career development.

Lisa Kozlowski, associate dean for postdoctoral affairs and recruitment at Thomas Jefferson University in Philadelphia, Pennsylvania, points out a less tangible achievement. "They've given postdocs a voice," she says, "and that's huge." ■

Karen Kaplan *is assistant Careers editor at* Nature.

1. Committee on Science, Engineering, and Public Policy *Enhancing the Postdoctoral Experience for Scientists and Engineers* (National Academies Press, 2000).
2. Committee on Bridges to Independence *Bridges to Independence* (National Academies Press, 2005).
3. Biomedical Research Workforce Working Group *Draft Report* (NIH, 2012).
4. Davis, G. *Am. Scientist* **93,** (3) Supplement http://postdoc.sigmaxi.org/results/ (2005).

TURNING POINT
David Shelly

David Shelly is a seismologist with the US Geological Survey (USGS) in Menlo Park, California. In December, he will receive the latest in a string of high-profile awards: the Macelwane Medal, presented at the American Geophysical Union conference in San Francisco, California.

D. SHELLY

You had a broad-based education at a liberal-arts college. How did that prepare you for earthquake research?
I studied maths and physics at Whitman College in Walla Walla, Washington, but knew that I wanted to do something more applied as a graduate student. I didn't have the background for a geology programme, so I studied geophysics at Stanford University in California. My broad liberal-arts background helped me to learn how to communicate ideas concisely: an important part of the scientific process.

Describe your PhD research.
I went to Tokyo to study subduction-zone tremors in 2005, as part of the East Asia and Pacific Summer Institutes funded by the US National Science Foundation and the Japan Society for the Promotion of Science. A typical earthquake is one sudden rupture of a fault, but low-frequency subduction-zone tremors are weak vibrations resulting from slow slip between tectonic-plate boundaries. They start and end gradually, yet last much longer than a normal earthquake. Such tremors are a challenge to work with: it is hard to distinguish the seismic-wave signals from the background noise.

Your work had a big impact on the field. Why?
After low-frequency events were discovered in 2002, it was unclear whether they were earthquake-like or more like volcanic tremors, with fluids moving below ground. I used a technique to identify low-frequency tremors without knowing the exact onset time of the wave phases, which overcame the signal-to-noise difficulties. My team's results suggested that low-frequency subduction-zone tremors can be generated by similar processes to, and on the same faults as, larger earthquakes (D. R. Shelly *et al. Nature* **446,** 305–307 (2007) and S. Ide *et al. Nature* **447,** 76–79; 2007). That got attention and was good for my career. I think not having preconceived ideas helped, as did being naive and willing to try untested approaches, and being one of the first people to work in the field.

How did the 2011 Japanese quake affect you?
It was shocking. I thought the early (conservative) reports of a magnitude-8.8 earthquake were a mistake. An hour later, I saw the tsunami footage and realized that this was a wake-up call about scenarios that could exist but haven't been observed for a few hundred years.

How did it affect earthquake research?
It raised the profile of earthquake forecasting in general. It was by far the best-recorded earthquake of that size ever. Having that gold mine of data is driving big parts of the field forward, and helps my group to maintain strong collaborative ties with Japan. There have been a lot of studies based on data from that event, and they will continue for decades.

In the course of your career, will scientists get closer to being able to predict earthquakes?
Some people think it is inherently impossible. I think there is a subset of earthquakes, like those triggered by a slow-slip event with tremors serving as an indicator, that can be predicted. Those that start small and cascade into a large event may be inherently unpredictable. Unfortunately, research funding remains a challenge overall.

What is it like to get so much recognition so early in your career?
It is flattering, and it is almost certainly good attention. I was shocked to get the Macelwane award; I didn't know that two colleagues at the USGS had nominated me. It is a lot to live up to. That said, it is good to have motivation for the future. I don't have any overarching research goals: I plan to keep my focus on the big picture, finding solutions to pressing problems. I also make sure I have a life outside science. Going camping and hiking helps me to avoid research burn-out. ■

INTERVIEW BY VIRGINIA GEWIN

WITHOUT

A powerful letter.

BY FRAN WILDE

I filled a glass of water before bed and that's when Tim finally shouted at me.

"Look at the calendar for Pete's sake, how many times do I have to tell you?" he said from the kitchen doorway.

I hadn't been home in eight months. Things were tense before I left. Now they were worse. The lag confused me, made it hard to remember what was restricted. Tim would be fined for my mistake; the seed corn, probably. *Mea culpa*.

I taught history. Resource allocation algorithms weren't my speciality. On the station, rations were calculated for us. Down here, folks tried to make restrictions easy to remember, to keep a kind of independence. A sense of choice. So, each restriction went with the day of the week: Wednesday, therefore water.

Dehydrated from the trip, I forgot. Might have been showing off a little, too.

Tomorrow would be better. Tomorrow was a T. Not too much started with T. Except toilet paper.

No, tomorrow wouldn't be good at all.

When I arrived early Wednesday morning, Tim welcomed me home. His face faltered when he saw I was alone. It broke my heart. I squeezed his hand, tried to make it up to him, to remind him what we once were. In the late afternoon, I woke from a siesta dream of ice in a glass, a bead of moisture clinging to the side, brushing my thumb, clinging. I heard laughter echoing down the halls of the dusty house. It evaporated when I woke, turned to motes of memory.

I couldn't leave until Tim signed the documents I'd brought, and he was stalling. Down here, he could stall me to death, if he wanted. At the very least, if he didn't sign by tomorrow, I'd be stuck until Saturday. No travel on Friday. Friday, therefore fuel. And food.

"Tim, this is awful. Why won't you emigrate?"

"Won't have to, with so many people leaving. Soon we won't need to restrict at all," he said.

He was stubborn, my husband. He had the whole homestead to himself now.

The cousins left first. They said to send word when things got better. Then his brothers went up.

They shut the school and offered me a job on the station. He stayed silent when he realized I was going. Ginned up a court order to speak for him. I had to sneak the girls out.

We watched from above while he held the fort, protected our heritage.

First thing when I landed, I showed him the photos of Joie and Darra. Thriving, I said. No dust to make them cough.

JACEY

But the station made it clear I couldn't keep them in my quarters without his signature, thanks to the court order. Couldn't enrol them in school, couldn't get them on the ration algorithm. I'd gone spare on my meals so they'd been eating fine. But I needed Tim to make it legal, or come up and be their father again.

I told him we had room for him. He set his face like a brick and turned to look at the land that ran right to the edge of the sea. It was brittle and dry. The wind blew hot. Last time he saw his daughters, their laughter rasped, mottled with dust and smoke.

"If I leave, it's gone forever. Without the land, there's nothing to come back to."

"What's to come back for?" I'd run my fingers through the scorched soil. It feathered like dust. "Won't be long before the cliff crumbles and the rest washes away."

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

"What if it doesn't?" He and the rest of the holdouts thought they could fix it.

"Sign the documents, Tim. Or come with me and do it yourself."

He didn't answer. He picked up the glass of water and looked at it in the light. I hadn't taken a sip. The area's cisterns must have run low. The water that came up through the homestead's pipes had dark tendrils of algae floating in it. First green thing I'd seen here, but not appealing.

He ran his thumb across the rim of the goblet, the etching. The glass was an old one. Mother called it 'depression glass' when she passed it down to me. I hadn't thought about how much of the past it contained when I held it under the tap. The light from the oil lamp reflected off the thick glass. The glitter ran its pattern across Tim's cheekbones, his jawline.

"It's important that someone stays," he said.

Life on the station wasn't perfect, but we had water every day. Days were just days. Green things grew. Filtered air flowed. There was a school. If he dropped the court order, I could sign the kids up for classes. Get them ready for the future.

"It's their lives, Tim."

"They won't remember living here."

"I'll make sure they will." I pulled out the documents again. The photos, too.

"Their future is here."

"Maybe someday. Until then, it's up there."

Other station residents were making the same plea across the scorched county. Mothers, fathers, children, grandchildren. One last try. Gravity tugged at our feet like it didn't want to let us go. The heat was beyond what we remembered.

"You come too," I said.

He shook his head. "I'll make do here." He signed without another word, then carried the glass out to the field and poured it over the dry earth. He didn't come back to the house.

I was free to go. He'd do without. Wednesday, therefore wife. ■

Fran Wilde *is a writer and technology consultant. She can also tie various sailing knots, set gemstones and program digital minions. She blogs at franwilde.wordpress.com.*